AD________________

Award Number: W81XWH-04-1-0495

TITLE: Investigation of Three-Group Classifiers to Fully Automate Detection and Classification of Breast Lesions in an Intelligent CAD Mammography Workstation

PRINCIPAL INVESTIGATOR: Darrin C. Edwards, Ph.D.
Charles E. Metz, Ph.D
Maryellen L. Giger, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, IL 60637

REPORT DATE: May 2007

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE 01-05-2007	**2. REPORT TYPE** Final	**3. DATES COVERED** 1 May 2004 – 30 Apr 2007

4. TITLE AND SUBTITLE

Investigation of Three-Group Classifiers to Fully Automate Detection and Classification of Breast Lesions in an Intelligent CAD Mammography Workstation

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0495

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Darrin C. Edwards, Ph.D., Charles E. Metz, Ph.D, Maryellen L. Giger, Ph.D.

Email: d-edwards@uchicago.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Chicago
Chicago, IL 60637

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our goal is to develop a fully automated classification scheme for computer-aided diagnosis in mammography. Our proposed scheme would classify computer detections into three groups: malignant lesions, benign lesions, and false-positive computer detections. We proved that the area under the ROC curve (AUC) is not useful in classification tasks with three or more groups, and showed that the three decision boundary lines used by the three-group ideal observer are intricately related to one another. We analyzed several recently proposed three-group classification methods in terms of the ideal observer. We collected a database of 270 mammographic images with clustered microcalcification lesions. We have developed a novel performance metric that may generalize better than AUC to tasks with more than two groups. A three-group classifier could potentially allow radiologists to detect more malignant breast lesions without increasing their false-positive biopsy rates.

15. SUBJECT TERMS
Computer-aided diagnosis, X-ray imaging

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	97	**19b. TELEPHONE NUMBER** *(include area code)*

1 Introduction 4

2 Body 4

3 Key Research Accomplishments 9

4 Reportable Outcomes 10

5 Conclusions 11

References 12

Appendices

A Bibliography 15

B List of Personnel 16

C The Hypervolume under the ROC Hypersurface of "Near-Guessing" and "Near-Perfect" Observers in N-Class Classification Tasks 17

D Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions 25

E Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule 35

F Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule 46

G Restrictions on the Three-Class Ideal Observer's Decision Boundary Lines 57

H Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities 66

I Optimization of restricted ROC surfaces in three-class classification tasks 74

J A utility-based performance metric for ROC analysis of N-class classification tasks 87

1 Introduction

Our goal is to develop a fully automated classification scheme for computer-aided diagnosis (CAD) in mammography. Traditional CAD classification schemes, and performance measurement tools such as receiver operating characteristic (ROC) analysis, are based on the premise that the observations are classified into two groups, most commonly malignant and benign. Such classification schemes are difficult to fully automate, as they analyze radiologist-identified lesions; this is because many false-positive (FP) detections produced by a computerized detection scheme cannot reasonably be classified as benign or malignant lesions. Our proposed scheme would classify computer detections into three groups: malignant lesions, benign lesions, and FP computer detections. This method presents considerable difficulties in terms of both signal detection theory and performance evaluation methods such as ROC analysis. Our efforts in this direction during the course of the supported research were thus generally more theoretical than practical. However, we consider the results of our work both promising and important.

2 Body

A wide variety of medical decision-making tasks, in particular tasks for which CAD has been proposed as an aid to the physician, can be formulated as “two-group classification” tasks. That is, the physician must use the information available about a patient (*e. g.*, a set of mammographic films of the patient, and the result of computer analysis of those images) to decide whether a patient belongs to a diseased, or abnormal, group or not (*e. g.*, whether a breast lesion suspicious enough to warrant further imaging procedures or biopsy is present or not).

ROC analysis has long been considered the most appropriate methodology for evaluating the performance of a two-group classifier or observer [1], particularly for medical decision-making tasks [2]. Furthermore, the optimal or “ideal” observer — that observer which achieves the best possible performance given a particular population of observational data — has also been well understood for quite some time [3]. In practice, the ideal observer requires knowledge of the probability density functions (PDFs) from which the observational data are drawn, and thus cannot be achieved in non-trivial tasks by human or automated observers. Nevertheless, successful methods for estimating ideal observer decision variables from a sample of observational data [4], and for plotting an ideal observer ROC curve from a sample of decision variable data [5], have been developed.

Although the form of the three-group ideal observer has also been known for some time [3], the development of a practical three-group classifier and a fully general extension of ROC analysis to three-group classification has proven quite difficult, primarily due to the tremendous increase in complexity encountered when one moves from two-group to three-group classification tasks. Briefly, characterizing the performance of a three-group classifier requires an ROC “hypersurface” with five degrees of freedom in a six-dimensional ROC space [6, 7] (by contrast, a two-group classifier is fully described by a simple curve in a two-dimensional ROC space). Despite these difficulties, our research efforts are focused on the development of a three-group classifier and performance evaluation methodology for breast lesion classification in a mammographic CAD system.

We strongly believe the development of such a three-group classifier to be of practical and not merely academic importance. In the past, two types of mammographic CAD schemes have been investigated at the University of Chicago: one for automatically detecting mass lesions in mammograms [8–12], and another for classifying known lesions as malignant or benign [13–17]. Combining these two types of CAD schemes is inherently difficult, because the output of the detection scheme, which identifies candidates for subsequent classification, will necessarily include FP computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our approach has been to treat this problem explicitly as a three-group classification task. That is, the output of the detection scheme should be classified as malignant lesions, benign lesions, and non-lesions (FP computer detections), and the classifier to be estimated is the ideal observer decision function for this task. If successful, this approach would allow radiologists to identify more malignant lesions without increasing biopsy rates for patients without malignancy.

Our approved Statement of Work was as follows:

Task 1. Develop a three-group classifier for clustered microcalcifications in mammograms, Months 1-12.

- (a) Collect cases containing 180 malignant and 180 benign clusters of microcalcifications.
- (b) Determine truth state of imaged lesions by reviewing the images, radiologist reports, and pathology reports for these cases.
- (c) Obtain at least 180 FP computer detections from these cases using the existing detection scheme.
- (d) Train and test a three-group classifier on these lesions, using methodology we previously developed for mass lesions.

Task 2. Design and develop an interface for an intelligent workstation for CAD, Months 11-14.

- (a) Examine the most useful features of the interface of the existing intelligent CAD workstation for mammographic lesion detection.
- (b) Examine the most useful features of the interface of the existing CAD schemes in our laboratory for classifying manually detected lesions as malignant or benign.
- (c) Develop a simple interface drawing on the advantages of the existing detection and classification schemes, extended to the three-group classification task.
- (d) Test the interface with non-radiologist observers in our laboratory familiar with the goals of CAD and with interface design principles.

Task 3. Design and perform a pilot observer study measuring radiologists' performances using the three-group classification schemes and traditional two-group classification schemes, Months 15-24.

- (a) Recruit radiologists from our institution and neighboring institutions.

(b) Provide training to the radiologists in the use of the intelligent CAD workstation interfaces.

(c) Measure radiologist performance using the three-group intelligent workstation, and using the existing intelligent workstation for detecting lesions followed by manual selection of lesions to be analyzed by the existing schemes for two-group classification of lesions.

Task 4. Develop techniques to compare radiologists' performance in using the proposed three-group and traditional two-group classification schemes, Months 18-36.

(a) Develop methodology to extend two-group ROC analysis to tasks in which observations are classified into three groups.

(b) Develop methodology to determine the statistical significance of measured differences in performance between three-group classifiers.

(c) Use this methodology to analyze the observer data obtained in Task 3.

For Tasks 1(a) and 1(b), we collected during the second year of this project a database of 134 mammographic cases, four standard views per case; the majority of these cases contained malignant or benign clustered microcalcification lesions. During the course of the past year, however, the images were found to be unsuitable for our purposes. We therefore collected another set of 270 images, 142 of which contained benign microcalcification clusters, and 128 of which contained malignant microcalcification clusters. The truth for the malignant microcalcification lesions was verified by pathology report, and that for the benign lesions by pathology report when biopsy was recommended, or by followup when that was recommended by the original radiologist. This is less than the number of malignant and benign lesions initially proposed for this project, but we will have the opportunity to supplement these with further such cases from the database of a colleague in our laboratories should the research continue under other funding mechanisms (see Sec. 4).

For Tasks 1(c) and 1(d), we initially encountered difficulties porting the computer code for the existing detection scheme from the legacy equipment for which it was written (IBM RISC 6000 machines, whose operating systems are no longer supported and whose hardware is too old to be considered reliable) to a modern PC workstation running a Linux operating system. These difficulties were traced to compiler incompatibilities between the two systems. A computer programmer in our laboratory with extensive experience with both systems and intimate familiarity with the internals of the detection scheme investigated and eliminated the majority of these.

We had planned to submit a paper to *Medical Physics* reporting on the results for Task 1. In fact, we are quite close to obtaining the final results needed for completing such a paper. Unfortunately, the principal investigator very recently discovered an error in the code he had written [18] to interface between the numerical programming environment we use (matlab) and the Bayesian artificial neural network (BANN) package of MacKay [19] that serves as the basis of our classifier [4,18]. We fully expect the relevant experiments incorporating the corrected code to be completed soon, and should be able to submit a paper describing these results to *Medical Physics* within another two or three weeks. We will then submit to the USAMRMC an addendum to this report including those final results.

Our research accomplishments focused largely on Task 4. Although the "methodology we previously developed for mass lesions" [20] was successful for estimating ideal observer decision *variables* based on lesion feature data, a practical classifier to make use of this decision variable data has not yet been implemented. As the difficulties in theoretically characterizing the behavior of such a three-group classifier are intimately related to evaluation of such a classifier's performance (*i. e.*, the development of a three-group extension to ROC analysis), such a reordering of the approved tasks seemed logically justified. In fact, the theoretical difficulties involved in completely characterizing the general behavior of a three-group ideal observer, and in developing a three-group extension to ROC analysis, prevented us from accomplishing Tasks 2 or 3. However, proposed further work on those theoretical issues, and on the development of such a classification scheme for CAD and its evaluation through radiologist observer studies, served as the basis for two research grant applications for which we have applied. These are listed in Sec. 4; if either is funded, it will provide support for the principal investigator at the assistant professor level.

By far the most important result achieved so far was our discovery and proof (published during the first year of support) that an obvious generalization of the well-known performance metric, the area under the ROC curve (AUC), is not in fact useful in tasks with three or more groups [21]. (See Appendix C.) This accomplishment relates directly to Task 4.(b) above, which implicitly requires a well-defined performance metric with respect to which the statistical significance of differences in performance may be computed. Although arguably a "negative" rather than "positive" result — a well-defined performance metric has not yet been found — this result has been very well received in the observer performance and CAD research communities. First, it serves as a striking yet typical example of how intuition can often be an unreliable guide in extending methodology from the two-group classification task to tasks with three or more groups. Second, it clearly indicates that the search for such a well-defined performance metric will yield a deeper understanding of the properties of three-group observer performance, particularly as characterized by ROC analysis.

We stated above that exact determination of the ideal observer's decision variables requires knowledge of the PDFs from which the observational data to be classified were drawn. The tool we have been using for some time now to estimate ideal observer decision variables from samples of observational data is the BANN [19]. In previous simulation studies in which the PDFs of the observational data are known, the output of the BANN was found to agree with the calculated ideal observer decision variables for two-group [4] and three-group [18] classification tasks. In practice, one does not have the PDFs of real observational data, but we previously developed a means of evaluating three-group BANN decision variables by comparing them with two-group BANN decision variables obtained from simplified two-group tasks using the same observational data [20]. During the first year of support, we developed an independent technique for evaluating three-group BANN estimates of ideal observer decision variables, again based on theoretical properties of the three-group ideal observer [22]. (See Appendix D.) This result is important because the three-group classifier we are developing under the current research will be trained and tested using feature data from actual mammograms; thus, we will not have access to the PDFs from which those data are drawn. In addition to three-group ROC analysis methods to be developed by extension from existing two-group methods [5], it will be beneficial to have a direct method of judging the ability of the BANN decision variables to accurately estimate ideal observer decision variables.

During the first and second years of support, we investigated in great detail the behavior of the three-group ideal observer. In particular, it is well-known that the three-group ideal observer makes decisions by partitioning a plane of two decision variables into three regions using three decision boundary lines [3]. We showed that the locations and orientations of these decision boundary lines are not arbitrary; given the slopes and y-intercepts, for example, of two of the lines, those of the third line are constrained to lie within a particular range of values [23]. (See Appendix G.) A detailed understanding of such properties of the three-group ideal observer will prove crucial to the calculation of observer ROC operating points, and by extension to observer performance evaluation in general.

In our efforts to develop a three-group classifier and appropriate performance evaluation methodology, we have made every attempt to keep our analysis as general as possible despite the theoretical difficulties this entails. Other researchers have proposed three-group methodology by considering observers whose behavior is restricted in particular ways, or by considering only a subset of the possible performance characterization indices (the axes of ROC space), or both [24–28]. The inherent complexity of the three-group classification task makes direct comparison of different methods by different researchers difficult. To facilitate such a comparison, we analyzed the different methods in terms of the three-group ideal observer, both in preliminary work [29] (see Appendix E) and later through more in-depth analysis [30]. (See Appendix F.) In addition to providing us with valuable insight and experience in comparing different classifiers, which should ultimately prove directly relevant to the completion of Task 4, this work also enabled us to present to the observer performance and CAD research communities a useful framework within which comparison of superficially very different classifiers can readily be made. A poster presentation of the theoretical results of this and the preceding paragraph, as well as our research accomplishments during the first year of this award, was made at the 2005 US DOD Breast Cancer Research Program Era of Hope Meeting in Philadelphia, PA [31].

In the second and third years of support, we analyzed a simplified performance evaluation method (*i. e.*, an extension of ROC analysis to tasks with three groups) which considers only the three "sensitivities" of the observer — the three probabilities of correctly identifying an observation from one of the three respective groups. (This can, in general, be expected to yield an incomplete description of observer performance, which requires a set of six conditional classification probabilities [7].) This method was originally proposed by Mossman [26] for a pair of essentially *ad hoc* decision rules and arbitrary decision variables, and more recently advocated by He *et al.* [28] for a set of ideal observer decision variables and a decision rule shown [28–30] to be a special case of the ideal observer decision rule, and also shown [29, 30] to be a special case of the decision rule proposed by Scurfield [25]. We were able to derive a more fundamental motivation for the decision rules described in those works, given the simplified performance description in terms of only the sensitivities, by applying previously successful Neyman-Pearson optimization methodology [3, 7] to this restricted performance evaluation strategy.

Simply put, assuming that one chooses to measure observer performance only in terms of the observer's sensitivities, we proved [32] that the optimal observer with respect to this metric is in fact the special case of the ideal observer proposed by He *et al.* [28]. (See Appendix H.) We then applied this analysis technique [33] to other decision strategies and performance evaluation strategies which we had previously analyzed in terms of the ideal observer decision rule [30]. (See Appendix I.) Given the difficulties inherent in a fully general description of three-class ideal observer behavior and performance evaluation, it is possible that a restricted or simplified model, similar to those proposed already by other researchers, may ultimately prove of greater practical value than the fully general theoretical model. We consider this work important, because it provides a principled theoretical framework in which to evaluate and compare such restricted and simplified models.

As stated above, a well-defined performance metric is required in order to understand the properties of three-group observer performance, particularly as characterized by ROC analysis. Furthermore, we showed that an obvious generalization of the AUC does not in fact prove useful in tasks with three or more groups [21]. During the third year of support, we developed, and presented preliminary results of studies involving, a novel "utility"-based performance metric [34]. (See Appendix J.) In the beginning of this section, we introduced the concept of the ideal observer as that observer which achieves the best possible performance given a particular population of observational data. One way of deriving the ideal observer model is to assign a number, the utility, to each possible decision; the ideal observer is then that observer which maximizes the expected utility [3, 7]. Our proposed performance metric is grounded in this concept of the utility of an observer's decisions, and can be shown to be directly related to intuitive properties of the observer's ROC curve (AUC and the arc length along the curve or, for tasks with more than two groups, the hypervolume under the ROC hypersurface and the hypersurface itself). Although further analysis will be necessary to fully characterize the properties of this novel performance metric, we have high hopes that it will prove to be of use in characterizing observer performance without being subject to the limitations we have shown exist for a more obvious generalization of the AUC.

A detailed understanding of the properties of the general three-group ideal observer, and of the restricted and simplified ROC models described above, will ultimately prove crucial to the calculation of observer ROC operating points, and by extension to observer performance evaluation in general. Throughout the course of this project, the principal investigator and mentor have held regular meetings to discuss the theoretical challenges posed by this project and to explore possible ways of overcoming those challenges.

3 Key Research Accomplishments

- Proof that an obvious generalization of the well-known two-group performance metric, the AUC, is not useful in classification tasks with three or more groups (Appendix C)
- Development of a novel technique for evaluating the quality of BANN estimates of ideal observer decision variables in the absence of three-group ROC analysis methodology and observational data PDFs (Appendix D)
- Detailed investigation of the relationships among the decision boundary lines used by the three-group ideal observer (Appendix G)
- Analysis of several proposed three-group classification methods in the literature in terms of the three-group ideal observer (Appendices E, F)
- Development of principled theoretical motivation for proposed three-group classification methods given selection of restricted or simplified three-group evaluation methodology (Appendices H, I)
- Development and preliminary analysis of a novel utility-based performance metric, which we hope will generalize better to classification tasks with more than two groups than does the conventional AUC (Appendix J)

4 Reportable Outcomes

- D. C. Edwards, C. E. Metz, and R. M. Nishikawa, “The hypervolume under the ROC hypersurface of ‘near-guessing’ and ‘near-perfect’ observers in N-class classification tasks,” *IEEE Trans. Med. Imag.*, vol. 24, pp. 293–299, 2005.

- D. C. Edwards and C. E. Metz, “Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions,” in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, Eds., SPIE, Bellingham, WA, 2005, pp. 174–182. [Conference presentation and proceedings paper.]

- D. C. Edwards and C. E. Metz, “Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule,” in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, Eds., SPIE, Bellingham, WA, 2005, pp. 128–137. [Conference presentation and proceedings paper.]

- Collection of database of 270 mammographic cases containing malignant and benign clustered microcalcification lesions, with truth determined by pathology (for biopsied lesions) or mammographic followup (benign lesions only)

- Porting of existing computerized scheme for detecting clustered microcalcifications in mammograms from legacy computer systems no longer in operation to workstations currently in use for this project

- D. C. Edwards, C. E. Metz, R. M. Nishikawa, and M. L. Giger, “Investigation of three-group classifiers to fully automate detection and classification of breast lesions in computer-aided diagnosis for mammography,” US DOD Breast Cancer Research Program Era of Hope Meeting, Philadelphia, PA, 2005.

- D. C. Edwards and C. E. Metz, “Restrictions on the three-class ideal observer’s decision boundary lines,” *IEEE Trans. Med. Imag.*, vol. 24, pp. 1566–1573, 2005.

- D. C. Edwards and C. E. Metz, “Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule,” *J. Math. Psychol.*, vol. 50, pp. 478–487, 2006.

- D. C. Edwards and C. E. Metz, “Optimization of an ROC hypersurface constructed only from an observer’s within-class sensitivities,” in Proc. SPIE Vol. 6146 *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Miguel P. Eckstein, Eds., SPIE, Bellingham, WA, 2006, pp. 61 460A1–61 460A7. [Conference presentation and proceedings paper.]

- D. C. Edwards and C. E. Metz, “ROC Analysis in Radiology: The State of the Art, and Recent N-Class Investigations,” Third Workshop on Receiver Operating Characteristic Analysis in Machine Learning, Pittsburgh, PA, 2006. (Invited talk.)

- D. C. Edwards and C. E. Metz, “A utility-based performance metric for ROC analysis of N-class classification tasks,” in Proc. SPIE Vol. 6515 *Medical Imaging 2007:*

Image Perception, Observer Performance, and Technology Assessment, Yulei Jiang and Berkman Sahiner, Eds., SPIE, Bellingham, WA, 2007, pp. 6 515 031–65 150 310. [Conference presentation and proceedings paper.]

- D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.*, 2006, (accepted for publication 5 Mar. 2007).
- D. C. Edwards, J. Papaioannou, C. E. Metz, A. V. Edwards, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic microcalcification lesions," 2007 (in preparation).
- D. C. Edwards, PI: "N-Class Image Classification for Computer-Aided Breast Cancer Diagnosis," application for support under NIH K99/R00 funding mechanism; submitted June 2006, unscored, resubmitted March 2007.
- D. C. Edwards, Project co-Leader under C. E. Metz (Project Leader) and R. M. Nishikawa (Program PI): "Three-class Receiver Operating Characteristic Analysis for Evaluation of Computer-Aided Diagnosis," Project 3 of Program Project Grant "Translating Computer-Aided Diagnosis (CADx) from the Lab to the Clinic," application for support under NIH P01 funding mechanism; submitted Oct. 2006, merit rating 1.7 (overall program priority score: 209), resubmitted May 2007.

5 Conclusions

During the first year of support, we proved that an obvious generalization of the well-known two-group performance metric, the AUC, is not in fact a useful performance metric for classification tasks with three or more groups. We developed an evaluation technique, independent of those we had previously developed, for assessing the ability of BANN decision variables to accurately estimate ideal observer decision variables. We analyzed several recently proposed three-group classification methods in terms of the three-group ideal observer. We also showed that the three decision boundary lines used by the three-group ideal observer are not arbitrary, but are intricately related to one another.

During the second year of support, with the assistance of colleagues in our laboratory, we collected a database of 134 mammographic cases containing malignant and benign clustered microcalcification lesions, with truth determined by pathology (for biopsied lesions) or mammographic followup (benign lesions only), and we ported the existing computerized scheme for detecting clustered microcalcifications in mammograms from legacy computer systems no longer in operation to workstations currently in use for this project. We reported on the important theoretical results we had developed to date at the 2005 Breast Cancer Research Program Era of Hope Meeting. We also developed principled theoretical motivations for various proposed three-group classification methods, given in each case the selection of a restricted or simplified three-group evaluation methodology.

Although the first set of images we collected proved unsuitable for our purposes, we were able during the past year to collect 270 mammographic images and are close to completing experiments, using these images, designed to evaluate the ability of BANNs to estimate ideal observer decision variables for mammographic lesion feature data (as opposed to simulated data). The principal investigator was invited to give a talk at the Third Workshop on

Receiver Operating Characteristic Analysis in Machine Learning (a conference within the International Conference on Machine Learning symposium) on the subject of the state of the art of ROC analysis in radiology and on our recent investigations in classification with more than two groups. We have also continued to advance our theoretical understanding of the three-group ideal observer and methods of evaluating its performance. In particular, we have developed a novel utility-based performance metric which we have reason to believe may be useful for classification tasks with more than two groups without suffering from the limitations of more obvious generalizations of the well-known AUC performance metric.

Although our primary research accomplishments have been theoretical, they are crucial steps in the development of a practical three-group classifier and a fully general three-group performance evaluation methodology. Despite the considerable difficulties involved in such development, a CAD scheme incorporating a three-group classifier as we propose could potentially allow radiologists to detect more malignant breast lesions without increasing their FP biopsy rate. We believe this goal to be worth the necessary effort on our part.

References

[1] J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.

[2] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine* **VIII**(4), pp. 283–298, 1978.

[3] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.

[4] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Trans. Med. Imag.* **20**, pp. 886–899, 2001.

[5] C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, pp. 1–33, 1999.

[6] C. Ferri, J. Hernández-Orallo, and M. A. Salido, "Volume under the ROC surface for multi-class problems: Exact computation and evaluation of approximations," tech. rep., Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València (Spain), 2003.

[7] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.

[8] U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.* **2**, pp. 1–9, 1995.

[9] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Med. Phys.* **18**, pp. 955–963, 1991.

[10] F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.* **28**, pp. 473–481, 1993.

[11] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.* **21**, pp. 445–452, 1994.

[12] M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging.* Ph.D. thesis, The University of Chicago, Chicago, IL, 2000.

[13] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.* **5**, pp. 155–168, 1998.

[14] Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.* **44**, pp. 2579–2595, 1999.

[15] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness," *Acad. Radiol.* **7**, pp. 1077–1084, 2000.

[16] Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.* **20**, pp. 1285–1292, 2001.

[17] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis — Observer study with independent database of mammograms," *Radiology* **224**, pp. 560–568, 2002.

[18] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network," in Proc. SPIE Vol. 4686 *Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, Dev P. Chakraborty and Elizabeth A. Krupinski, eds., pp. 1–12, (SPIE, Bellingham, WA), 2002.

[19] D. J. S. MacKay, *Bayesian Methods for Adaptive Models.* Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.

[20] D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.* **31**, pp. 81–90, 2004.

[21] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks," *IEEE Trans. Med. Imag.* **24**, pp. 293–299, 2005.

[22] D. C. Edwards and C. E. Metz, "Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, eds., pp. 174–182, (SPIE, Bellingham, WA), 2005.

[23] D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.* **24**, pp. 1566–1573, 2005.

[24] B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math. Psychol.* **40**, pp. 253–269, 1996.

[25] B. K. Scurfield, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math. Psychol.* **42**, pp. 5–31, 1998.

[26] D. Mossman, "Three-way ROCs," *Med. Decis. Making* **19**, pp. 78–89, 1999.

[27] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 567–578, (SPIE, Bellingham, WA), 2003.

[28] X. He, C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis — A decision theoretic approach under the ideal observer framework," *IEEE Trans. Med. Imag.* **25**, pp. 571–581, 2006.

[29] D. C. Edwards and C. E. Metz, "Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, eds., pp. 128–137, (SPIE, Bellingham, WA), 2005.

[30] D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.* **50**, pp. 478–487, 2006.

[31] D. C. Edwards, C. E. Metz, R. M. Nishikawa, and M. L. Giger, "Investigation of three-group classifiers to fully automate detection and classification of breast lesions in computer-aided diagnosis for mammography." US DOD Breast Cancer Research Program Era of Hope Meeting, Philadelphia, PA, 2005.

[32] D. C. Edwards and C. E. Metz, "Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities," in Proc. SPIE Vol. 6146 *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Miguel P. Eckstein, eds., pp. 61460A1–61460A7, (SPIE, Bellingham, WA), 2006.

[33] D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.* , 2006. (accepted for publication 5 Mar 2007).

[34] D. C. Edwards and C. E. Metz, "A utility-based performance metric for ROC analysis of N-class classification tasks," in Proc. SPIE Vol. 6515 *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Berkman Sahiner, eds., pp. 6515031–65150310, (SPIE, Bellingham, WA), 2007.

A Bibliography

Listed below are all publications and meeting abstracts which were supported by this award.

- D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks," *IEEE Trans. Med. Imag.*, vol. 24, pp. 293–299, 2005.

- D. C. Edwards and C. E. Metz, "Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, Eds., SPIE, Bellingham, WA, 2005, pp. 174–182. [Conference presentation and proceedings paper.]

- D. C. Edwards and C. E. Metz, "Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, Eds., SPIE, Bellingham, WA, 2005, pp. 128–137. [Conference presentation and proceedings paper.]

- D. C. Edwards, C. E. Metz, R. M. Nishikawa, and M. L. Giger, "Investigation of three-group classifiers to fully automate detection and classification of breast lesions in computer-aided diagnosis for mammography," US DOD Breast Cancer Research Program Era of Hope Meeting, Philadelphia, PA, 2005.

- D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.*, vol. 24, pp. 1566–1573, 2005.

- D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.*, vol. 50, pp. 478–487, 2006.

- D. C. Edwards and C. E. Metz, "Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities," in Proc. SPIE Vol. 6146 *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Miguel P. Eckstein, Eds., SPIE, Bellingham, WA, 2006, pp. 61 460A1–61 460A7. [Conference presentation and proceedings paper.]

- D. C. Edwards and C. E. Metz, "ROC Analysis in Radiology: The State of the Art, and Recent N-Class Investigations," Third Workshop on Receiver Operating Characteristic Analysis in Machine Learning, Pittsburgh, PA, 2006. (Invited talk.)

- D. C. Edwards and C. E. Metz, "A utility-based performance metric for ROC analysis of N-class classification tasks," in Proc. SPIE Vol. 6515 *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Berkman Sahiner, Eds., SPIE, Bellingham, WA, 2007, pp. 6 515 031–65 150 310. [Conference presentation and proceedings paper.]

- D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.*, 2006, (accepted for publication 5 Mar. 2007).

B List of Personnel

Listed below are all personnel who were supported by this award.

- Darrin C. Edwards, principal investigator
- Charles E. Metz, mentor
- Maryellen L. Giger, scientific collaborator
- Lorenzo Pesce, computer programmer

C The Hypervolume under the ROC Hypersurface of "Near-Guessing" and "Near-Perfect" Observers in N-Class Classification Tasks

The Hypervolume Under the ROC Hypersurface of "Near-Guessing" and "Near-Perfect" Observers in N-Class Classification Tasks

Darrin C. Edwards*, Charles E. Metz, and Robert M. Nishikawa

***Abstract*—We express the performance of the N-class "guessing" observer in terms of the $N^2 - N$ conditional probabilities which make up an N-class receiver operating characteristic (ROC) space, in a formulation in which sensitivities are eliminated in constructing the ROC space (equivalent to using false-negative fraction and false-positive fraction in a two-class task). We then show that the "guessing" observer's performance in terms of these conditional probabilities is completely described by a degenerate hypersurface with only $N - 1$ degrees of freedom (as opposed to the $N^2 - N - 1$ required, in general, to achieve a true hypersurface in such a ROC space). It readily follows that the hypervolume under such a degenerate hypersurface must be zero when $N > 2$. We then consider a "near-guessing" task; that is, a task in which the N underlying data probability density functions (pdfs) are nearly identical, controlled by $N - 1$ parameters which may vary continuously to zero (at which point the pdfs become identical). With this approach, we show that the hypervolume under the ROC hypersurface of an observer in an N-class classification task tends continuously to zero as the underlying data pdfs converge continuously to identity (a "guessing" task). The hypervolume under the ROC hypersurface of a "perfect" ideal observer (in a task in which the N data pdfs never overlap) is also found to be zero in the ROC space formulation under consideration. This suggests that hypervolume may not be a useful performance metric in N-class classification tasks for $N > 2$, despite the utility of the area under the ROC curve for two-class tasks.**

***Index Terms*—N-class classification, ROC analysis, ROC performance metrics.**

I. Introduction

We are attempting to develop a fully automated mass lesion classification scheme for computer-aided diagnosis (CAD) in mammography. This scheme will combine two schemes developed at the University of Chicago: one for automatically detecting mass lesions in mammograms [1]–[5], and one for classifying known lesions as malignant or benign [6]–[10]. Combining these two types of CAD scheme is inherently difficult, because the output of the detection scheme will necessarily include false-positive (FP) computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our approach has been to treat this problem explicitly as a three-class classification task. That is, the outputs of the detection scheme should be classified as malignant lesions, benign lesions, and nonlesions (FP computer detections), and the classifier to be estimated is the ideal observer decision function for this task. Such an approach presents considerable difficulties of its own. On the one hand, decision functions, in particular ideal observer decision functions, increase rapidly in complexity with the number of classes involved. On the other hand, fully general performance evaluation methods, in particular a fully general three-class extension of receiver operating characteristic (ROC) analysis, have yet to be developed for such a task.

Although we have had preliminary success in using Bayesian artificial neural networks (BANNs) [11], [12] to estimate three-class ideal-observer-related decision variables [13], [14], the task of developing an extension of ROC analysis to classification tasks with three or more classes has proved somewhat more daunting. Our initial efforts in this direction have, thus, been more theoretical than practical so far [15]. One issue we began to investigate recently was the calculation of an obvious generalization of the well-known area under the ROC curve (AUC) performance metric, a quantity we are calling the "hypervolume under the ROC hypersurface." Detailed consideration of the integrals involved in calculating this quantity led us to the counterintuitive conclusion that, despite the great success and utility of the AUC performance metric in two-class classification tasks, the hypervolume under the ROC hypersurface does not appear to be a useful performance metric in N-class classification tasks for $N > 2$. The proof of this claim is arrived at by considering observer performance in two extremes: the "guessing" observer and the "perfect" observer. It should be explicitly noted that in our formulation, sensitivities are eliminated in constructing the ROC space; this is equivalent to using false-negative fraction (FNF) and false-positive fraction (FPF) in a two-class task. In such a formulation, the "guessing" observer in a two-class task achieves an AUC of 0.5 as expected, but the "perfect" observer in a two-class task achieves an AUC of zero.

In Section II, we consider the properties of the "guessing" observer in an N-class classification task, and of its ROC

Manuscript received June 23, 2004; revised November 8, 2004. This work was supported in part by the National Cancer Institute under Grant R01-CA60187 and in part by the US Army Breast Cancer Research Program under Grant W81XWH-04-1-0495. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was E. Krupinski. *Asterisk indicates corresponding author.*

*D. C. Edwards is with the Department of Radiology, The University of Chicago, 5841 So. Maryland Avenue, Chicago, IL 60637 USA (e-mail: d-edwards@uchicago.edu)

C. E. Metz and R. M. Nishikawa are with the Department of Radiology, The University of Chicago, Chicago, IL 60037 USA.

Digital Object Identifier 10.1109/TMI.2004.841227

hypersurface. In Section III, we consider the properties of the ROC hypersurface of a so-called "near-guessing" observer, i.e., an observer in a task for which the observational data probability density functions (pdfs) are not identical, but differ only by arbitrarily small amounts. In Section IV, we then show that the hypervolume under the ROC hypersurface of such a "near-guessing" observer will continuously approach the hypervolume under the ROC hypersurface of the "guessing" observer as the observational data pdfs continuously approach identity; furthermore, the hypervolume under the ROC hypersurface of the "guessing" observer is shown to be zero.

We then show in Section V that the hypervolume under the ROC hypersurface of the "perfect" observer is zero (as expected by analogy with the two-class task), and that the hypervolume under the ROC hypersurface of a "near-perfect" observer will approach zero continuously as the observational data pdfs are separated. Finally, in Section VI, we argue that these results taken together imply that the hypervolume under the ROC hypersurface is not a useful performance metric in N-class classification tasks for $N > 2$, despite the utility of the AUC performance metric in two-class tasks.

II. The ROC Hypersurface of the N-Class "Guessing" Observer

The performance of an observer in an N-class classification task is completely determined by a hypersurface with $N^2 - N - 1$ degrees of freedom in an $(N^2 - N)$-dimensional ROC space [16]. Without loss of generality, we can specify any point in the ROC space by a vector of the misclassification probabilities $[P(\mathbf{d} = \pi_1 \,|\, \mathbf{t} = \pi_2), \ldots, P(\mathbf{d} = \pi_1 \,|\, \mathbf{t} = \pi_N), P(\mathbf{d} = \pi_2 \,|\, \mathbf{t} = \pi_1), P(\mathbf{d} = \pi_2 \,|\, \mathbf{t} = \pi_3), \ldots, P(\mathbf{d} = \pi_2 \,|\, \mathbf{t} = \pi_N), P(\mathbf{d} = \pi_N \,|\, \mathbf{t} = \pi_{N-1}), P(\mathbf{d} = \pi_N \,|\, \mathbf{t} = \pi_1)]^\dagger$ [15]. Here the N classes are denoted by the labels $\pi_1, \ldots, \pi_N$; $\mathbf{d}$ denotes the class to which an observation is assigned (the "decision"); and $\mathbf{t}$ is the class to which it actually belongs (the "truth"). We use boldface type to denote statistically variable quantities. For simplicity, we write $P(\mathbf{d} = \pi_i \,|\, \mathbf{t} = \pi_j)$ as P_{ij}.

We can also, again without loss of generality, consider the ROC hypersurface to be given by P_{N1} considered as a function of the other $N^2 - N - 1$ misclassification probabilities [15]. Note that this formulation is equivalent, in a two-class classification task, to using FPF and FNF to characterize the ROC curve, rather than FPF and true-positive fraction (TPF), as is more common. In a two-class classification task, this produces ROC curves which are "upside-down" with respect to the standard formulation; we have adopted the nonstandard formulation described above because it has proven easier to generalize to classification tasks with more than two classes.

Some researchers have suggested [17], [18] that in, e.g., a three-class classification task, the set of three "sensitivities" ($P(\mathbf{d} = \pi_i \,|\, \mathbf{t} = \pi_i)$ in our notation) provides a complete description of observer performance. This is incorrect in general, because it ignores the $N^2 - N$ misclassification probabilities, not all of which are determined uniquely by the "sensitivities" when $N > 2$ unless particular restrictions are imposed on the observer's behavior. Complete quantification of the trade-offs available among the probabilities of various kinds of misclassification error is important in medical diagnosis, where different misclassification errors often have substantially different clinical consequences. Moreover, restrictions concerning the observer's behavior are inappropriate when considering the general behavior of ideal observers, human observers, or automated observers (such as automated schemes for computer-aided diagnosis) designed to approximate ideal or human observer behavior. Other researchers have reduced the three-class ROC hypersurface to more tractable two-dimensional surfaces in three-dimensional ROC spaces by explicitly imposing restrictions on the form of the observer's decision rule [19], [20], or on the utilities used by an ideal observer [21]. While such restrictions may ultimately prove to be of great pragmatic importance given the inherent complexity of multi-class classification tasks, our approach so far has been to attempt as general an understanding as possible of the unrestricted classification task.

Consider the performance of an observer which makes decisions by "guessing," that is, in a random fashion unrelated to the actual class $\mathbf{t}$ from which a given observation is drawn. (Note that this corresponds to the performance of the ideal observer when the pdfs of the observational data are identical, i.e., $p(\vec{\mathbf{x}} \,|\, \pi_1) = p(\vec{\mathbf{x}} \,|\, \pi_2) = \cdots = p(\vec{\mathbf{x}} \,|\, \pi_N)$.) In this case, we clearly must have

$$P_{12} = P_{13} = \cdots = P_{1N} \tag{1}$$

$$P_{21} = P_{23} = \cdots = P_{2N} \tag{2}$$

$$\cdots$$

$$P_{N1} = P_{N2} = \cdots = P_{N\,(N-1)}. \tag{3}$$

Defining $\alpha_i \equiv P_{iN}$ for $1 \le i \le N-1$, and $\alpha_N \equiv P_{N\,(N-1)}$, we see that the performance of the "guessing" observer is given by a locus of vectors of the form

$$\begin{bmatrix} \left.\begin{matrix}\alpha_1\\ \alpha_1\\ \vdots\end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix}\alpha_i\\ \alpha_i\\ \vdots\end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix}\alpha_N\\ \alpha_N\\ \vdots\end{matrix}\right\} N-1 \text{ elements} \end{bmatrix} \tag{4}$$

where all of the α_i are restricted to the range $[0, 1]$. Furthermore, note that

$$\begin{aligned} P(\mathbf{d} = \pi_i) &= \sum_{j=1}^{N} P_{ij} P(\mathbf{t} = \pi_j) \\ &= \sum_{j=1}^{N} \alpha_i \, P(\mathbf{t} = \pi_j) \\ &= \alpha_i \end{aligned} \tag{5}$$

which immediately gives $\alpha_N = 1 - \sum_{i=1}^{N-1} \alpha_i$. Thus, the performance of the “guessing” observer is given by

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{N\,(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \left.\begin{matrix} \alpha_1 \\ \alpha_1 \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} \alpha_i \\ \alpha_i \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} 1 - \sum_{j=1}^{N-1} \alpha_j \\ 1 - \sum_{j=1}^{N-1} \alpha_j \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \end{bmatrix}$$

$$= \vec{v}_0 + \sum_{i=1}^{N-1} \alpha_i \vec{v}_i. \tag{6}$$

This is the parametric equation for an $(N-1)$-dimensional plane in an $(N^2 - N)$-dimensional space; the actual performance of the “guessing” observer will of course be further restricted to a region within this plane such that $0 \leq \alpha_i \leq 1, 0 \leq 1 - \sum \alpha_i \leq 1$.

III. The ROC Hypersurface of an N-Class “Near-Guessing” Observer

Consider observational data $\vec{\mathbf{x}}$ drawn from N pdfs

$$p(\vec{x}\,|\,\mathbf{t} = \pi_1) = p(\vec{x}\,|\,\mathbf{t} = \pi_N) + \delta_1 h_1(\vec{x}) \tag{7}$$

$$\cdots$$

$$p(\vec{x}\,|\,\mathbf{t} = \pi_j) = p(\vec{x}\,|\,\mathbf{t} = \pi_N) + \delta_j h_j(\vec{x}) \tag{8}$$

$$\cdots$$

$$p(\vec{x}\,|\,\mathbf{t} = \pi_N) \tag{9}$$

where $0 \leq \delta_j \leq 1, \int h_j(\vec{x}) d^n\vec{x} = 0$, and $|h_j(\vec{x})| \leq p(\vec{x}\,|\,\mathbf{t} = \pi_N)$ for $1 \leq j \leq N-1$. In the limit as the δ_j all approach zero, we expect the performance of any observer for this task to converge smoothly to that of the “guessing” observer.

Decisions are made by partitioning the decision variable space into N regions, determined by a total of $N^2 - N - 1$ parameters; we denote these parameters by the components of a vector $\vec{\gamma}$. An observer which uses more than $N^2 - N - 1$ parameters for an N-class classification task can always be replaced by a simplified observer, such that the “excess” parameters are eliminated by the requirement that P_{N1} be minimized, thereby collapsing the dimensionality of the parameter space to $N^2 - N - 1$. On the other hand, an observer which uses fewer than $N^2 - N - 1$ decision parameters will fail to generate a true ROC hypersurface—i.e., one with $N^2 - N - 1$ degrees of freedom in the $(N^2 - N)$-dimensional ROC space. (An example in a three-class classification task would be an observer which sequentially performs a pair of binary classification tasks by first classifying observations as being “π_1” or “not π_1” based on the value of a single decision parameter, and then further classifying the “not π_1” observations as “π_2” or “π_3” based on the value of a second decision parameter [17], thus depending on fewer than the five degrees of freedom needed in a three-class classification task.) Such “degenerate” observers will not be considered here (apart from the “guessing” observer itself).

We can, thus, define N regions which partition the original data space, given particular values of the parameters $\vec{\gamma}$, by

$$\mathcal{D}_1(\vec{\gamma}) \equiv \{\vec{x} : \mathbf{d} = \pi_1 \text{ given } \vec{\gamma}\} \tag{10}$$

$$\cdots$$

$$\mathcal{D}_i(\vec{\gamma}) \equiv \{\vec{x} : \mathbf{d} = \pi_i \text{ given } \vec{\gamma}\} \tag{11}$$

$$\cdots$$

$$\mathcal{D}_N(\vec{\gamma}) \equiv \{\vec{x} : \mathbf{d} = \pi_N \text{ given } \vec{\gamma}\}. \tag{12}$$

For a nonrandom observer, the $\mathcal{D}_i$ can be expected to depend implicitly on the pdfs (7)–(9) and, therefore, on the δ_j. The misclassification probabilities which define the ROC hypersurface are then given by

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{N\,(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \int_{\mathcal{D}_1} p(\vec{x}\,|\,\mathbf{t} = \pi_2)\, d^n\vec{x} \\ \int_{\mathcal{D}_1} p(\vec{x}\,|\,\mathbf{t} = \pi_3)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_1} p(\vec{x}\,|\,\mathbf{t} = \pi_N)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_i} p(\vec{x}\,|\,\mathbf{t} = \pi_1)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_i} p(\vec{x}\,|\,\mathbf{t} = \pi_j)\, d^n\vec{x} \quad \{i \neq j\} \\ \vdots \\ \int_{\mathcal{D}_i} p(\vec{x}\,|\,\mathbf{t} = \pi_N)\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_N} p(\vec{x}\,|\,\mathbf{t} = \pi_{N-1})\, d^n\vec{x} \\ \vdots \\ \int_{\mathcal{D}_N} p(\vec{x}\,|\,\mathbf{t} = \pi_1)\, d^n\vec{x} \end{bmatrix}. \tag{13}$$

Using (7) and (8), we can rewrite this as

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{N\,(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} P_{1N} + \delta_2 \int_{\mathcal{D}_1} h_2(\vec{x})\, d^n\vec{x} \\ P_{1N} + \delta_3 \int_{\mathcal{D}_1} h_3(\vec{x})\, d^n\vec{x} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{iN} + \delta_1 \int_{\mathcal{D}_i} h_1(\vec{x})\, d^n\vec{x} \\ \vdots \\ P_{iN} + \delta_j \int_{\mathcal{D}_i} h_j(\vec{x})\, d^n\vec{x} \; \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{N\,N} + \delta_{N-1} \int_{\mathcal{D}_N} h_{N-1}(\vec{x})\, d^n\vec{x} \\ \vdots \\ P_{N\,N} + \delta_1 \int_{\mathcal{D}_N} h_1(\vec{x})\, d^n\vec{x} \end{bmatrix}. \tag{14}$$

Defining the functions $H_{ij} \equiv \int_{\mathcal{D}_i} h_j(\vec{x})\, d^n\vec{x}$ allows us to simplify the notation slightly

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{(N-1} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} P_{1N} + \delta_2 H_{12} \\ P_{1N} + \delta_3 H_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{iN} + \delta_1 H_{i1} \\ \vdots \\ P_{iN} + \delta_j H_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{N\,N} + \delta_{N-1} H_{N(N-1)} \\ \vdots \\ P_{N\,N} + \delta_1 H_{N1} \end{bmatrix}. \tag{15}$$

Now of course $P_{N\,N} = 1 - \sum_{i=1}^{N-1} P_{iN}$; for simplicity, we will write $\alpha_i \equiv P_{iN}$. Equation (15) can now be written as

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \alpha_1 + \delta_2 H_{12} \\ \alpha_1 + \delta_3 H_{13} \\ \vdots \\ \alpha_1 \\ \vdots \\ \alpha_i + \delta_1 H_{i1} \\ \vdots \\ \alpha_i + \delta_j H_{ij} \quad \{i \neq j\} \\ \vdots \\ \alpha_i \\ \vdots \\ 1 - \sum_{j=1}^{N-1} \alpha_j + \delta_{N-1} H_{N(N-1)} \\ \vdots \\ 1 - \sum_{j=1}^{N-1} \alpha_j + \delta_1 H_{N1} \end{bmatrix} \tag{16}$$

which further simplifies to

$$\begin{bmatrix} P_{12} \\ P_{13} \\ \vdots \\ P_{1N} \\ \vdots \\ P_{i1} \\ \vdots \\ P_{ij} \quad \{i \neq j\} \\ \vdots \\ P_{iN} \\ \vdots \\ P_{(N-1)} \\ \vdots \\ P_{N1} \end{bmatrix} = \begin{bmatrix} \left.\begin{matrix} \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} \alpha_i \\ \alpha_i \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \\ \vdots \\ \left.\begin{matrix} 1 - \sum_{j=1}^{N-1} \alpha_j \\ 1 - \sum_{j=1}^{N-1} \alpha_j \\ \vdots \end{matrix}\right\} N-1 \text{ elements} \end{bmatrix} + \sum_{j=1}^{N-1} \delta_j \vec{w}_j \tag{17}$$

where the vectors $\vec{w}_j$ have components which depend only on H_{ij}. The first term on the right-hand side of this equation is just the expression for the "guessing" observer [cf. the left-hand side of (6)]. The other term on the righthand side of this equation tends to zero as the δ_j tend to zero. Note that the H_{ij} may in general depend on the δ_k via (10)–(12), but

$$\begin{aligned} |H_{ij}| &= \left| \int_{\mathcal{D}_i} h_j(\vec{x})\, d^n\vec{x} \right| \\ &\leq \int_{\mathcal{D}_i} |h_j(\vec{x})|\, d^n\vec{x} \\ &\leq \int_{\mathcal{D}_i} p(\vec{x}\,|\,\mathbf{t} = \pi_N)\, d^n\vec{x} \\ &= P_{iN} \\ &\leq 1. \end{aligned} \tag{18}$$

Thus, the H_{ij} are bounded, and will possess Taylor expansions in δ_k (i.e., will not depend on terms of the form δ_k^{-m} for positive integers m). Therefore, operating points on the ROC hypersurface of a "near-guessing" observer tend continuously toward points on the ROC hypersurface of the "guessing" observer. Note that the $N(N-1)$ terms $\alpha_i, \delta_j H_{ij}$ are not all independent, since they all depend implicitly for fixed δ_j on the $N^2 - N - 1$ decision parameters $\vec{\gamma}$. That is, the ROC hypersurface given by (17) possesses only $N^2 - N - 1$ degrees of freedom.

IV. The Hypervolume Under the ROC Hypersurface of an N-Class "Near-Guessing" Observer

In the preceding section, it was shown that the ROC hypersurface of a "near-guessing" observer tends continuously to the ROC hypersurface of a "guessing" observer as the pdfs of the observational data tend arbitrarily toward identical distributions. Intuitively, one would expect that the hypervolumes under these hypersurfaces should also tend toward each other. Since intuition can occasionally be an unreliable guide in analyzing N-class classification tasks, it would be reassuring if the results of the preceding section could be applied directly to the calculation of the relevant hypervolumes.

For this section, we will write P_{ij} as $P_{ij}(\vec{\gamma})$, emphasizing that it is a function of the decision parameters chosen. We, thus, rewrite (15) to obtain

$$\begin{bmatrix} P_{12}(\vec{\gamma}) \\ P_{13}(\vec{\gamma}) \\ \vdots \\ P_{1N}(\vec{\gamma}) \\ \vdots \\ P_{i1}(\vec{\gamma}) \\ \vdots \\ P_{ij}(\vec{\gamma}) \quad \{i \neq j\} \\ \vdots \\ P_{iN}(\vec{\gamma}) \\ \vdots \\ P_{N,(N-1)}(\vec{\gamma}) \\ \vdots \\ P_{N,1}(\vec{\gamma}) \end{bmatrix} = \begin{bmatrix} P_{1N}(\vec{\gamma}) + \delta_2 H_{12}(\vec{\gamma}) \\ P_{1N}(\vec{\gamma}) + \delta_3 H_{13}(\vec{\gamma}) \\ \vdots \\ P_{1N}(\vec{\gamma}) \\ \vdots \\ P_{iN}(\vec{\gamma}) + \delta_1 H_{i1}(\vec{\gamma}) \\ \vdots \\ P_{iN}(\vec{\gamma}) + \delta_j H_{ij}(\vec{\gamma}) \quad \{i \neq j\} \\ \vdots \\ P_{iN}(\vec{\gamma}) \\ \vdots \\ P_{N,(N-1)}(\vec{\gamma}) \\ \vdots \\ P_{N,(N-1)}(\vec{\gamma}) - \delta_{N-1} H_{N(N-1)}(\vec{\gamma}) \\ +\delta_1 H_{N1}(\vec{\gamma}) \end{bmatrix}. \tag{19}$$

To find the hypervolume under the ROC surface given by P_{N1} considered as a function of $(P_{12}, P_{13}, \ldots, P_{ij}, \ldots, P_{N(N-1)}, \ldots, P_{N2})$, one must evaluate the integral

$$\int \cdots \int P_{N1} d^{N^2-N-1}\vec{P}. \tag{20}$$

(The domain of the integral is simply the set of all P_{ij} such that P_{N1} is defined.) Note that, for the "guessing" observer, we expect this integral to be zero when $N > 2$ due to dimensionality considerations—the ROC hypersurface has only $N-1$ degrees of freedom (cf. (6)), not the N^2-N-1 required in this (N^2-N)-dimensional ROC space. To see this explicitly, one can rearrange the order of integration and consider the innermost integral $\int P_{N1} dP_{N(N-1)}$ for fixed values of the other misclassification probabilities. Then the limits of integration of this innermost definite integral become, again by (6)

$$\int_{1-\sum P_{jN}}^{1-\sum P_{jN}} P_{N1} dP_{N(N-1)} \quad \{j < N\} \tag{21}$$

which is zero by inspection.

We now return to the general case of a "near-guessing" observer. One way to evaluate the integral in (20) is to reexpress it explicitly in terms of the decision parameters $\vec{\gamma}$, via the Jacobian

$$J \equiv \begin{vmatrix} \frac{\partial P_{12}}{\partial \gamma_1} & \frac{\partial P_{12}}{\partial \gamma_2} & \frac{\partial P_{12}}{\partial \gamma_3} & \cdots & \frac{\partial P_{12}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{1N}}{\partial \gamma_1} & \frac{\partial P_{1N}}{\partial \gamma_2} & \frac{\partial P_{1N}}{\partial \gamma_3} & \cdots & \frac{\partial P_{1N}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{ij}}{\partial \gamma_1} & \frac{\partial P_{ij}}{\partial \gamma_2} & \frac{\partial P_{ij}}{\partial \gamma_3} & \cdots & \frac{\partial P_{ij}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N(N-1)}}{\partial \gamma_1} & \frac{\partial P_{N(N-1)}}{\partial \gamma_2} & \frac{\partial P_{N(N-1)}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N2}}{\partial \gamma_1} & \frac{\partial P_{N2}}{\partial \gamma_2} & \frac{\partial P_{N2}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N2}}{\partial \gamma_{N^2-N-1}} \end{vmatrix} \tag{22}$$

where the vertical bars indicate that the determinant of the enclosed matrix is to be taken, and where γ_i denotes the ith component of $\vec{\gamma}$. (We assume that indices of the parameters $\vec{\gamma}$ have been chosen appropriately so that no negative sign is introduced, i.e., volumes remain positive.) For the "guessing" observer, this reduces to

$$\begin{aligned} &J_{\text{guessing}} \\ &\equiv \begin{vmatrix} \frac{\partial P_{1N}}{\partial \gamma_1} & \frac{\partial P_{1N}}{\partial \gamma_2} & \frac{\partial P_{1N}}{\partial \gamma_3} & \cdots & \frac{\partial P_{1N}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{1N}}{\partial \gamma_1} & \frac{\partial P_{1N}}{\partial \gamma_2} & \frac{\partial P_{1N}}{\partial \gamma_3} & \cdots & \frac{\partial P_{1N}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{iN}}{\partial \gamma_1} & \frac{\partial P_{iN}}{\partial \gamma_2} & \frac{\partial P_{iN}}{\partial \gamma_3} & \cdots & \frac{\partial P_{iN}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N(N-1)}}{\partial \gamma_1} & \frac{\partial P_{N(N-1)}}{\partial \gamma_2} & \frac{\partial P_{N(N-1)}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_{N^2-N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_{N(N-1)}}{\partial \gamma_1} & \frac{\partial P_{N(N-1)}}{\partial \gamma_2} & \frac{\partial P_{N(N-1)}}{\partial \gamma_3} & \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_{N^2-N-1}} \end{vmatrix} \end{aligned} \tag{23}$$

where $P_{N(N-1)} = P_{NN} = 1 - \sum_{j=1}^{N-1} P_{jN}$. For a "near-guessing" observer, we combine (19) and (22) to obtain

$$J_{\text{near}} \equiv \begin{vmatrix} \cdots & \frac{\partial(P_{1N}+\delta_2 H_{12})}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial P_{1N}}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial(P_{iN}+\delta_j H_{ij})}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial P_{N(N-1)}}{\partial \gamma_k} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \frac{\partial(P_{N(N-1)}-\delta_{N-1} H_{N(N-1)}+\delta_2 H_{N2})}{\partial \gamma_k} & \cdots \end{vmatrix}. \tag{24}$$

From the properties of determinants [22], it can be shown that, to first order in the δ_j,

$$J_{\text{near}} = J_{\text{guessing}} + \sum_{j=1}^{N-1} \delta_j J_j + \cdots \tag{25}$$

where the J_j are bounded and continuous with respect to the δ_j.

If we denote the hypervolume under the ROC hypersurface of the "guessing" observer by

$$\begin{aligned} I_{\text{guessing}} &= \int \cdots \int P_{N1}\, d^{N^2-N-1}\vec{P} \\ &= \int \cdots \int P_{N(N-1)}(\vec{\gamma}) J_{\text{guessing}}\, d^{N^2-N-1}\vec{\gamma} \end{aligned} \tag{26}$$

then the hypervolume under the ROC hypersurface of a "near-guessing" observer becomes, again to first order in the δ_j

$$\begin{aligned} I_{\text{near}} = \int \cdots \int &[P_{N,(N-1)}(\vec{\gamma}) - \delta_{N-1} H_{N(N-1)}(\vec{\gamma}) \\ &+ \delta_1 H_{N1}(\vec{\gamma})] \\ &\times \left[J_{\text{guessing}} + \sum_{j=1}^{N-1} \delta_j J_j + \cdots \right] d^{N^2-N-1}\vec{\gamma} \end{aligned} \tag{27}$$

$$= I_{\text{guessing}} + \sum_{j=1}^{N-1} \delta_j I_j + \cdots \tag{28}$$

where the integrals I_j are bounded (i.e., they may depend on higher integral powers of δ_j, but not on δ_j^{-m} for positive integers m). That is, in the limit as the δ_j tend toward zero, I_{near} tends toward I_{guessing} in a continuous fashion.

V. The Hypervolume Under the ROC Hypersurface of an N-Class "Near-Perfect" Observer

In the preceding sections, we established that the hypervolume under the ROC hypersurface of a "guessing" observer is zero, and furthermore that this result is not singular: an observer in a "near-guessing" task will achieve a ROC hypersurface with hypervolume approaching zero continuously as the data pdfs approach identity. An ideal observer in a "perfect" task—i.e., in which the data pdfs never overlap—will also achieve a ROC hypersurface with zero hypervolume, because it can achieve the operating point $\vec{0}$ and, thus, will not, for any rational decision rule, achieve points interior to the unit hypercube defining ROC space. It is reasonable to ask whether "near-perfect" observers, performing tasks for which the overlap in the underlying data pdfs is nearly negligible, behave similarly to "near-guessing" observers, in the sense that the hypervolume under the ROC hypersurface of such an observer will approach zero in a continuous fashion.

Consider observational data $\vec{\mathbf{x}}$ drawn from N pdfs $p(\vec{x}\,|\,\mathbf{t} = \pi_j)$ where $1 \leq j \leq N$. We denote the mean of $p(\vec{x}\,|\,\mathbf{t} = \pi_j)$ by $\vec{\mu}_j$ and note that, without loss of generality, the mean of $p(\vec{x}\,|\,\mathbf{t} = \pi_N)$ can be taken to be $\vec{0}$. Furthermore, note that we can apply a linear transformation to the data $\vec{\mathbf{x}}$ and, thus, effectively to the $\vec{\mu}_j$, such that each of the resulting $\vec{\mu}_j$ is either 1) mutually orthogonal to, or 2) a scalar multiple of, any of the other $\vec{\mu}_i$. Because the transformation applied is linear, the ideal observer for this task will remain the same, and hence the task itself can be considered essentially unchanged.

Let us consider now an observer for this task which is generally not ideal; in fact, we will consider only a single operating point achieved by this observer. The observer decides $d = \pi_i$ for a given observation $\vec{x}$ if

$$(\vec{x} - \vec{\mu}_i) \cdot \frac{(\vec{\mu}_j - \vec{\mu}_i)}{|\vec{\mu}_j - \vec{\mu}_i|} < \frac{1}{2}|\vec{\mu}_j - \vec{\mu}_i| \qquad \{j : 1 \leq j \leq N, j \neq i\} \tag{29}$$

with equality for any such relation between two classes being decided in an arbitrary but consistent manner. That is, the observer places hyperplanes between the means of any two classes when attempting to decide between those classes (rather than placing those hyperplanes in the likelihood ratio decision variable space, as would the ideal observer).

Now suppose the task is made slightly "easier," while the observer itself remains unchanged. That is, consider the mean of one pdf, say $\vec{\mu}_i$ for $i \neq N$, being increased by a factor $1 + \delta$ for $0 \leq \delta \leq 1$, while the location of the decision hyperplanes does not change, except in the special case where $\vec{\mu}_j = \alpha\vec{\mu}_i$ for some other pdf (again with $j \neq N$). In this latter case we increase both means $(\vec{\mu}'_j = (1 + \delta)\vec{\mu}_j, \vec{\mu}'_i = (1 + \delta)\vec{\mu}_i)$, and the location of the corresponding decision hyperplane shifts accordingly.

Note that $\vec{\mu}'_i$ is now further away from each decision hyperplane relevant to $\mathbf{d} = \pi_i$ in (29). In the case $\vec{\mu}_j = \alpha\vec{\mu}_i$, the decision hyperplane is now a distance of $|(\vec{\mu}'_j) - (\vec{\mu}'_i)/(2)| = (1+\delta)|(\vec{\mu}_j) - (\vec{\mu}_i)/(2)|$ from $\vec{\mu}'_i$. For noncollinear $\vec{\mu}_j$, the direction from $\vec{\mu}'_i$ to the decision hyperplane is given by $\vec{\mu}_j - \vec{\mu}_i$, and since $\vec{\mu}_j$ and $\vec{\mu}'_i$ are orthogonal, $(\vec{\mu}'_i - \vec{\mu}_i) \cdot (\vec{\mu}_j - \vec{\mu}_i) = -\delta|\vec{\mu}_i|^2$; since this quantity is negative, it follows that $\vec{\mu}'_i$ is further from that decision plane than $\vec{\mu}_i$.

It immediately follows from this that none of the misclassification probabilities making up the coordinates of the observer's operating point can increase when moving from the old task to the new one. To see this, consider a change of coordinates in the data space such that $\vec{\mu}'_i$ is now the origin. All of the decision hyperplanes separating this class from the others are effectively moving away from the center of its pdf; since the hyperplanes are translating without rotating, we see immediately that the probability P_{ii} cannot decrease (and will increase in general), while the other probabilities P_{ji} $(j \neq i)$ cannot increase (and will decrease in general).

Note that any pdf $p(\vec{x})$ must decrease more rapidly than $|\vec{x}|^{-n}$ for sufficiently large $|\vec{x}|$, where n is the dimensionality of $\vec{x}$. This allows us to state qualitatively the sense in which the observer under consideration is "near-perfect": we hypothesize that the $|\vec{\mu}_i|$ are all sufficiently large that this limiting condition is met. Given this condition, the only situation in which an error probability P_{ji} $(j \neq i)$ will fail to decrease is if this probability is already zero. By allowing all of the $|\vec{\mu}_i|$ to increase in the manner described above, we can clearly obtain in general a situation in which each of the misclassification probabilities is either decreasing, or equal to zero.

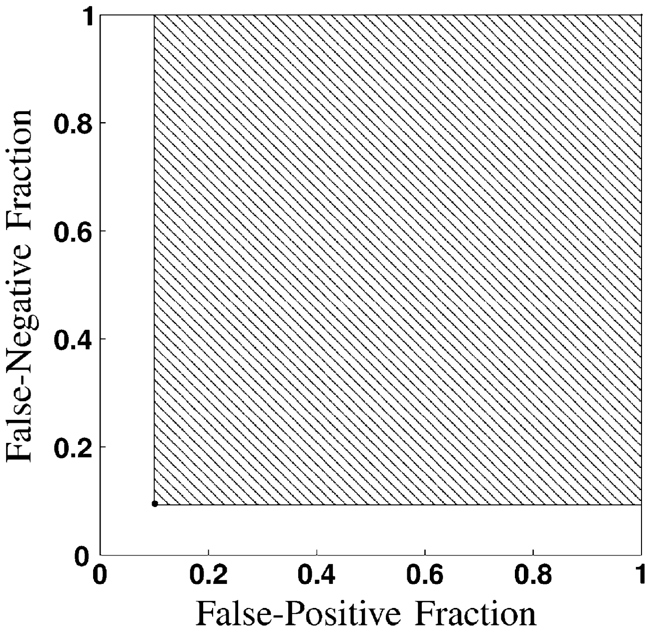


Fig. 1. Operating point of an observer in a two-class classification task with coordinates $(\mathrm{FPF}_0, \mathrm{FNF}_0)$, denoted by the point at the lower left corner of the crosshatched region. Since no rational observer will achieve points in the crosshatched region, the area under this observer's ROC curve cannot be greater than $1 - (1 - \mathrm{FPF}_0)(1 - \mathrm{FNF}_0)$.

This implies that the hypervolume under the ROC hypersurfaces of the observers under consideration (however we chose to define their decision rules for operating points other than those described above) must also decrease as the task is made "easier" as described above. To see this, note that if a given observer achieves an operating point $\vec{P}$ on its ROC hypersurface, it cannot achieve another point $\vec{P}'$ such that the components of these points satisfy $P'_i > P_i (1 \leq i \leq N^2 - N)$ (because such an observer could be replaced by an observer which achieved $\vec{P}$ for all such points by using the original decision rule for the point $\vec{P}$, thereby achieving unambiguously better performance at those points). Thus, knowing that a given observer achieves an operating point of $\vec{P}$ implies that that observer's ROC hypersurface must have a hypervolume under it of no greater than $1 - \prod_{i=1}^{N^2-N}(1 - P_i)$; as the (nonzero) P_i decrease, this upper limit on the hypervolume must also decrease to zero. This point is illustrated in Fig. 1 for the two-class case; here the observer's false-negative fraction, FNF_0, corresponds to P_{21}, and the false-positive fraction, FPF_0, corresponds to P_{12}.

To summarize, we have shown that the known operating point of our simple observer will move closer to the origin for arbitrary data pdfs as those pdfs are moved further apart (i.e., as the underlying task is made "easier"), implying that the hypervolume under its ROC hypersurface will also converge to zero. In fact, reasoning as above, one can see that the ideal observer will also be unable to achieve operating points within the region $P_i' > P_i$ $(1 \leq i \leq N^2 - N)$, since the ideal observer's ROC hypersurface is never above that of any other observer at any given point in the domain of the ROC space [15]. The hypervolume under the ideal observer's ROC hypersurface will, thus, also converge to zero as the underlying data pdfs are moved apart.

VI. Conclusion

In N-class classification tasks where $N > 2$, it can be shown that the hypervolume under the ROC hypersurface of both the "guessing" observer and the "perfect" observer are zero. More importantly, we have shown in each of these performance extremes that the convergence to zero is smooth rather than discontinuous. This convergence can be considered completely general for "near-guessing" observers and generally true for "near-perfect" observers which follow rational decision rules (analogous to false-negative fraction and false-positive fraction being monotonically related in a two-class task); that is, the conclusions appear to hold true for arbitrary underlying data pdfs.

In the two-class classification task, the area under the ROC curve (AUC) is considered a useful performance metric for a variety of reasons. One of the most pleasing and straightforward of these is the simple relationship between AUC and the "separability" of the two underlying data pdfs (i.e., the difficulty of the task). Namely, the AUC (with the two-class ROC defined as a plot of false-negative fraction versus false-positive fraction) of a "perfect" observer is zero, and increases in some sense uniformly as the task is made more difficult, until one arrives at the "guessing" observer with an AUC of 0.5. In an N-class classification task, this straightforward relationship appears to break down, and both "perfect" and "guessing" observers yield ROC hypersurfaces with zero hypervolume. It would appear that, due to this ambiguity, hypervolume under the ROC hypersurface of an N-class observer is not a useful performance metric: Does a hypervolume of 0.005 indicate an observer faced with an exceptionally difficult or exceptionally easy task? One hopes that some other performance metric from two-class classification can be generalized usefully for N-class classification; perhaps a quantity which is equal to AUC in the two-class case has a generalization which is not equal to the hypervolume, but can be shown to be of use for other reasons.

Acknowledgment

The authors thank the associate editor and anonymous reviewers for their helpful suggestions for improving the quality of this manuscript.

References

[1] U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.*, vol. 2, pp. 1–9, 1995.

[2] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Med. Phys.*, vol. 18, pp. 955–963, 1991.

[3] F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.*, vol. 28, pp. 473–481, 1993.

[4] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.*, vol. 21, pp. 445–452, 1994.

[5] M. A. Kupinski, "Computerized Pattern Classification in Medical Imaging," Ph.D. Thesis, The University of Chicago, Chicago, IL, 2000.

[6] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[7] Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.*, vol. 44, pp. 2579–2595, 1999.

[8] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness," *Acad. Radiol.*, vol. 7, pp. 1077–1084, 2000.

[9] Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.*, vol. 20, pp. 1285–1292, 2001.

[10] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis—Observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560–568, 2002.

[11] D. J. S. MacKay, "Bayesian Methods for Adaptive Models," Ph.D. Thesis, California Institute of Technology, Pasadena, CA, 1992.

[12] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Trans. Med. Imag.*, vol. 20, pp. 886–899, 2001.

[13] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network," in *Proc. SPIE, Vol. 4686, Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, D. P. Chakraborty and E. A. Krupinski, Eds. Bellingham, WA, 2002, pp. 1–12.

[14] D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.*, vol. 31, pp. 81–90, 2004.

[15] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.*, vol. 23, pp. 891–895, Jul. 2004.

[16] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Part I*. New York: Wiley, 1968.

[17] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, pp. 78–89, 1999.

[18] S. Dreiseitl, L. Ohno-Machado, and M. Binder, "Comparing three-class diagnostic tests by three-way ROC analysis," *Med. Decis. Making*, vol. 20, pp. 323–331, 2000.

[19] B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math Psychol.*, vol. 40, pp. 253–269, 1996.

[20] ——, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psychol.*, vol. 42, pp. 5–31, 1998.

[21] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study," in *Proc. SPIE, Vol. 5032, Medical Imaging 2003: Image Processing*, vol. 5032, M. Sonka and J. M. Fitzpatrick, Eds. Bellingham, WA, 2003, pp. 567–578.

[22] S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations*, 2nd ed. San Diego, CA: Harcourt Brace Jovanovich, 1986.

D Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions

Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions

Darrin C. Edwards* and Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

Bayesian artificial neural networks (BANNs) have proven useful in two-class classification tasks, and are claimed to provide good estimates of ideal-observer-related decision variables (the *a posteriori* class membership probabilities). We wish to apply the BANN methodology to three-class classification tasks for computer-aided diagnosis, but we currently lack a fully general extension of two-class receiver operating characteristic (ROC) analysis to objectively evaluate three-class BANN performance. It is well known that "the likelihood ratio of the likelihood ratio is the likelihood ratio." Based on this, we found that the decision variable which is the *a posteriori* class membership probability of an observational data vector is in fact equal to the *a posteriori* class membership probability of that decision variable. Under the assumption that a BANN can provide good estimates of these *a posteriori* probabilities, a second BANN trained on the output of such a BANN should perform very similarly to an identity function. We performed a two-class and a three-class simulation study to test this hypothesis. The mean squared error (deviation from an identity function) of a two-class BANN was found to be 2.5×10^{-4}. The mean squared error of the first component of the output of a three-class BANN was found to be 2.8×10^{-4}, and that of its second component was found to be 3.8×10^{-4}. Although we currently lack a fully general method to objectively evaluate performance in a three-class classification task, circumstantial evidence suggests that two- and three-class BANNs can provide good estimates of ideal-observer-related decision variables.

Keywords: Bayesian artificial neural networks, ideal observers, three-class classification

1. INTRODUCTION

In the past, computerized methods for the detection[1–5] and classification[6–11] of mammographic mass lesions have been investigated at the University of Chicago. The classification scheme currently analyzes lesions which have been manually identified by a radiologist. We are attempting to develop a fully automated classification scheme by combining the existing detection and classification schemes; we have argued previously[12] that this will require a three-class classifier to account for the presence of false-positive (FP) computer detections, in addition to the malignant and benign lesions, in the output of the detection scheme.

For some time now we have explored the use of Bayesian artificial neural networks (BANNs) for a variety of detection[5, 13, 14] and classification[11] tasks in computer-aided diagnosis (CAD). Our motivation for investigating BANNs is based, first, on our theoretical observation that, in the limit of infinite training data, a BANN will yield an ideal observer decision function for that data population;[15] and second, on empirical observations that even given a finite sample of training data, a BANN can estimate an ideal observer decision function reasonably well.[16] (We note that the BANN implementation we are using is that of MacKay,[17] which employs a multivariate normal function for the prior distribution on the network weight values.) We have also performed simulation studies showing that BANNs can accurately estimate ideal observer decision variables in a three-class classification task.[15] Moreover, we showed recently that a three-class BANN could produce decision variables for actual mammographic mass lesion feature data, and that these decision variables are related to two-class BANN decision variable data in a particular way consistent with a theoretical relationship between three-class and two-class ideal observer decision variables.[12] We consider this to be strong circumstantial evidence for the ability of a BANN to estimate three-class ideal observer decision variables, though we currently lack a fully general method for evaluating three-class classifiers (*i.e.*, a three-class extension to receiver operating characteristic (ROC) analysis).

*Correspondence: E-mail: d-edwards@uchicago.edu; Telephone: 773 834 5094; Fax: 773 702 0371

In this work, we present further circumstantial evidence toward the claim that a BANN can provide good estimates of three-class ideal observer decision variables. We develop a theoretical relationship between the *a posteriori* class membership probabilities of a given observational data variable and the *a posteriori* class membership probabilities of those *a posteriori* probabilities treated as a set of observational data in their own right. (It is known that *a posteriori* class membership probabilities are equivalent to ideal observer decision variables in a two-class task,[16] and related in a straightforward way to the ideal observer decision variables in a task with three or more classes.[15]) We then describe simulation studies to train and test a set of BANNs, and present results of such a simulation study verifying that the BANNs we examined did indeed obey the theoretical relationship predicted for ideal observer decision variables, to within experimental error. In the final section, we present our conclusions drawn from this work.

2. THEORY

It is well known that the ideal observer decision variable, *i.e.*, the likelihood ratio or any monotonic transformation of this value, yields optimal performance in a two-class classification task.[18] It can also be shown, in a classification task with N classes ($N > 2$), that the ideal observer decision rule becomes more complicated than a simple threshold on a single decision variable, but that the optimal decision variables remain a set of $N - 1$ likelihood ratios.[18, 19]

We can define the ith likelihood ratio as

$$\boldsymbol{\Lambda}_i \equiv \mathrm{LR}_i(\vec{\mathbf{x}}) \equiv \frac{p(\vec{\mathbf{x}}|\pi_i)}{p(\vec{\mathbf{x}}|\pi_N)}, \tag{1}$$

where $\vec{\mathbf{x}}$ represents statistically variable observational data (which we assume to have dimensionality n), and π_j represents one of the N classes from which the data are drawn (here $1 \leq i \leq N - 1$). Clearly the vector (of dimensionality $N - 1$) of decision variables $\boldsymbol{\Lambda}_i$ is itself statistically variable, and one might ask what the likelihood ratios of these variables are. In fact,[20]

$$\begin{aligned} p(\vec{\Lambda}|\pi_i) &= \int\cdots\int \sum_j \frac{p(\vec{x}_j|\pi_i)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \int\cdots\int \sum_j \mathrm{LR}_i(\vec{x}_j) \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n, \end{aligned} \tag{2}$$

where we have assumed that $N - 1 < n$; if $N - 1 = n$, then no integration is performed. (If $N - 1 > n$, then at least one of the likelihood ratio decision variables will be expressible as a function of the others; we will not consider this degenerate case here.) The sum is over all solutions to Eq. 1 for a given $\vec{\Lambda}$; this yields

$$\begin{aligned} p(\vec{\Lambda}|\pi_i) &= \int\cdots\int \sum_j \Lambda_i \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \Lambda_i \int\cdots\int \sum_j \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \Lambda_i\ p(\vec{\Lambda}|\pi_N) \\ \frac{p(\vec{\Lambda}|\pi_i)}{p(\vec{\Lambda}|\pi_N)} &\equiv \mathrm{LR}_i(\vec{\Lambda}) = \Lambda_i, \end{aligned} \tag{3}$$

the source of the well-known adage that "the likelihood ratio of the likelihood ratio is the likelihood ratio."

Consider now a different set of decision variables, the *a posteriori* class membership probabilities considered as functions of the statistically variable observational data

$$\mathbf{y}_i \equiv P(\pi_i|\vec{\mathbf{x}}). \tag{4}$$

(Since $P(\pi_N|\vec{x}) = 1 - \sum_{i=1}^{N-1} P(\pi_i|\vec{x})$, we still have $N-1$ decision variables.) Note that in a two-class classification task, this decision variable is known to be a monotonic function of the likelihood ratio, and is therefore an ideal observer decision variable;[16] while in a classification task with more than two classes, the *a posteriori* class membership probabilities can be shown to be related to the likelihood ratios in a straightforward way.[15]

Reasoning as above, we may ask what the *a posteriori* class membership probability of these decision variables, or $P(\pi_i|\vec{\mathbf{y}})$, is. In fact,

$$\begin{aligned} P(\pi_i|\vec{x}) &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{p(\vec{x})} \\ &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{\sum_{k=1}^{N} p(\vec{x}|\pi_k)P(\pi_k)} \\ &= \frac{\mathrm{LR}_i(\vec{x})P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \mathrm{LR}_k(\vec{x})P(\pi_k)/P(\pi_N)}, \qquad (5) \end{aligned}$$

and this relation can also be inverted to yield

$$\begin{aligned} \mathrm{LR}_i(\vec{x}) &= \frac{P(\pi_i|\vec{x})}{1 - \sum_{k=1}^{N-1} P(\pi_k|\vec{x})P(\pi_k)/P(\pi_N)} \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)}. \qquad (6) \end{aligned}$$

We again start with Eq. 2, this time obtaining

$$\begin{aligned} p(\vec{y}|\pi_i) &= \int \cdots \int \sum_j \mathrm{LR}_i(\vec{x}_j) \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \int \cdots \int \sum_j \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} \int \cdots \int \sum_j \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|}\, dx^N \ldots dx^n \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} p(\vec{y}|\pi_N), \qquad (7) \end{aligned}$$

where the sums in j are over all solutions to Eq. 4 for a given $\vec{y}$. (The fraction can be taken out of the integral because the relations in Eqs. 5 and 6 are one-to-one, and thus the set of all solutions to Eq. 4 correspond to a single value of $\mathrm{LR}_i(\vec{x_j})$.) This again yields

$$\mathrm{LR}_i(\vec{y}) = \mathrm{LR}_i(\vec{x}_j) \qquad (8)$$

where $\vec{y}$ is the vector of *a posteriori* class membership probabilities of $\vec{x}$ from Eq. 4, and $\vec{x}_j$ is any solution to that equation for a given $\vec{y}$.

It follows that

$$\begin{aligned} P(\pi_i|\vec{y}) &= \frac{\mathrm{LR}_i(\vec{y})P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \mathrm{LR}_k(\vec{y})P(\pi_k)/P(\pi_N)} \\ &= \frac{\mathrm{LR}_i(\vec{x}_j)P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \mathrm{LR}_k(\vec{x}_j)P(\pi_k)/P(\pi_N)} \\ &= P(\pi_i|\vec{x}_j) = y_i, \qquad (9) \end{aligned}$$

where $\vec{x}_j$ is again any solution to Eq. 4 for a given $\vec{y}$. This shows that a similar adage to that for likelihood ratios holds true, namely that "the *a posteriori* class probabilities of the (data) *a posteriori* class probabilities are the (data) *a posteriori* class probabilities."

3. MATERIALS AND METHOD

We have shown in the past[16] that a BANN can provide good estimates of the *a posteriori* class membership probabilities in a two-class classification task, and we have presented the results of simulation studies[15] and experiments with real mammographic feature data[12] strongly suggesting that the same holds true for three-class BANNs as well. The theoretical relationship given by Eq. 9, derived in the preceding section, provides a basis for another simulation study which should provide further circumstantial evidence for the claim that two-class and three-class BANNs can provide good estimates of the two- and three-class *a posteriori* class membership probabilities (directly related to the ideal observer decision variables *via* Eq. 5), respectively.

Specifically, for the two-class simulation study, we drew 500 samples pseudorandomly from each of two distributions:

$$p(x|\pi_1) \equiv N(x;\mu_1 = 1, \sigma_1^2 = 2) \tag{10}$$

$$p(x|\pi_2) \equiv N(x;\mu_2 = 0, \sigma_2^2 = 1). \tag{11}$$

We then trained a two-class BANN with one input, five hidden units, and one output on this data, obtaining a classifier we denote by

$$y = B_1^2(x). \tag{12}$$

(The superscript denotes the number of classes being classified.) We then used this output, given the known truth states for the original observations x from which it was obtained, as training data for a second BANN with one input, five hidden units, and one output:

$$z = B_2^2(y). \tag{13}$$

Finally, we pseudorandomly sampled an independent testing set of 500 observations x from each of the two classes given in Eqs. 10 and 11. This testing set was used as input to the first BANN to obtain a testing set y^{test}; this in turn was given as input to the second BANN, for which the output was z^{test}.

Given Eq. 9, together with the assumption that an adequately trained two-class BANN yields good estimates of the *a posteriori* class membership probabilities of the observations being classified, it should be the case that z^{test} estimates y^{test} at least to within experimental error. To verify this, we plotted z^{test} as a function of y^{test} for each of the two classes, and we computed the mean squared error

$$\text{MSE}_2 = \frac{1}{1000}\sum (z^{\text{test}} - y^{\text{test}})^2, \tag{14}$$

where the sum is over all the observations in the two classes.

Similarly, for the three-class simulation study, we drew 500 two-dimensional samples pseudorandomly from each of three distributions:

$$p(\vec{x}|\pi_1) \equiv N\left(\vec{x};\vec{\mu}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 4 & .75 \times 2 \\ .75 \times 2 & 1 \end{bmatrix}\right) \tag{15}$$

$$p(\vec{x}|\pi_2) \equiv N\left(\vec{x};\vec{\mu}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & -.4 \times 1.5 \\ -.4 \times 1.5 & 2.25 \end{bmatrix}\right) \tag{16}$$

$$p(\vec{x}|\pi_3) \equiv N\left(\vec{x};\vec{\mu}_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{17}$$

We then trained a three-class BANN with two inputs, five hidden units, and two outputs on this data, obtaining a classifier we denote by

$$\vec{y} = B_1^3(\vec{x}). \tag{18}$$

We then used this output, given the known truth states for the original observations $\vec{x}$ from which it was obtained, as training data for a second BANN with two inputs, five hidden units, and two outputs:

$$\vec{z} = B_2^3(\vec{y}). \tag{19}$$

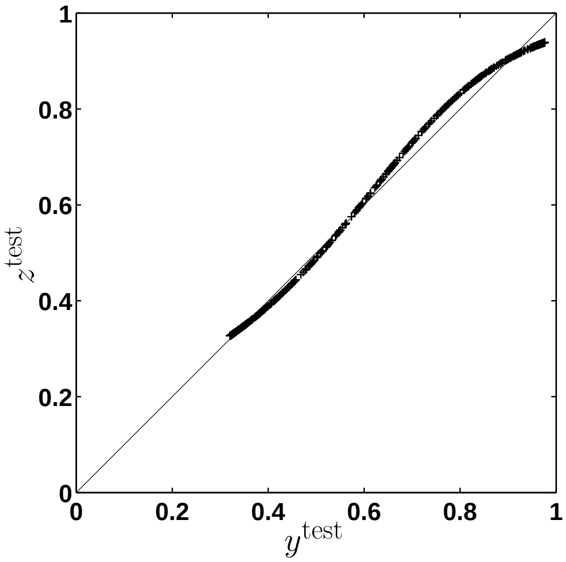


Figure 1. Output of the second two-class BANN as a function of its input for the observations actually drawn from class π_1 in the two-class simulation study.

Finally, we pseudorandomly sampled an independent testing set of 500 observations $\vec{x}$ from each of the three classes given in Eqs. 15-17. This testing set was used as input to the first BANN to obtain a testing set $\vec{y}^{\text{ test}}$; this in turn was given as input to the second BANN, for which the output was $\vec{z}^{\text{ test}}$.

Again, given Eq. 9, together with the assumption that an adequately trained two-class BANN yields good estimates of the *a posteriori* class membership probabilities of the observations being classified, it should be the case that z_1^{test} estimates y_1^{test}, and z_2^{test} estimates y_2^{test}, at least to within experimental error. To verify this, we plotted z_1^{test} as a function of y_1^{test}, and z_2^{test} as a function of y_2^{test}, for each of the three classes, and we computed the mean squared errors

$$\text{MSE}_{3i} = \frac{1}{1500}\sum(z_i^{\text{test}} - y_i^{\text{test}})^2, \tag{20}$$

$\{i : 1, 2\}$, where the sum is over all the observations in the three classes.

4. RESULTS

Figure 1 shows z^{test} as a function of y^{test} for the observations in class π_1, and Fig. 2 shows z^{test} as a function of y^{test} for the observations in class π_2 from the two-class simulation study. The mean squared error for the complete set of 1000 observations was 2.5×10^{-4}.

Figure 3 shows the components of $\vec{z}^{\text{ test}}$ as a function of the corresponding components of $\vec{y}^{\text{ test}}$ for the observations in class π_1. Similarly Fig. 4 shows the components of $\vec{z}^{\text{ test}}$ as a function of the corresponding components of $\vec{y}^{\text{ test}}$ for the observations in class π_2, and Fig. 5 shows the components of $\vec{z}^{\text{ test}}$ as a function of the corresponding components of $\vec{y}^{\text{ test}}$ for the observations in class π_3. The mean squared error for the complete set of 1500 observations was 2.8×10^{-4} for the first component and 3.8×10^{-4} for the second component.

5. DISCUSSION AND CONCLUSIONS

We developed a theoretical relationship between the *a posteriori* class membership probabilities, directly related to ideal observer decision variables, and the *a posteriori* class membership probabilities of those *a posteriori* class membership probabilities treated as statistically variable observer data in their own right. The identity relationship found is, perhaps unsurprisingly, quite similar in spirit to the identity relationship between the likelihood ratio decision variables and the likelihood ratio of those likelihood ratio decision variables for a given task.

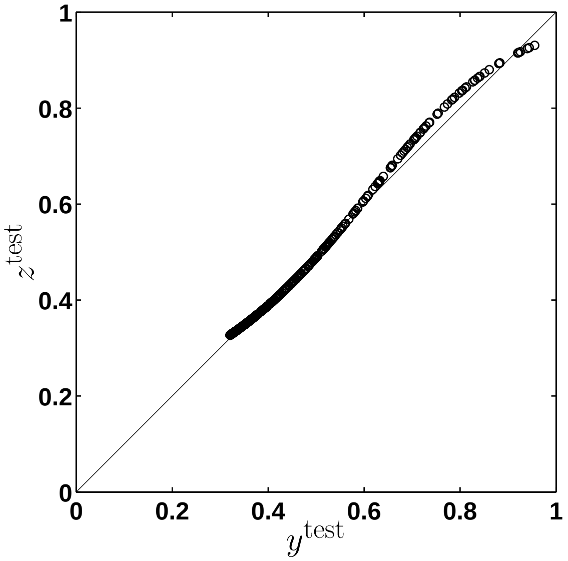
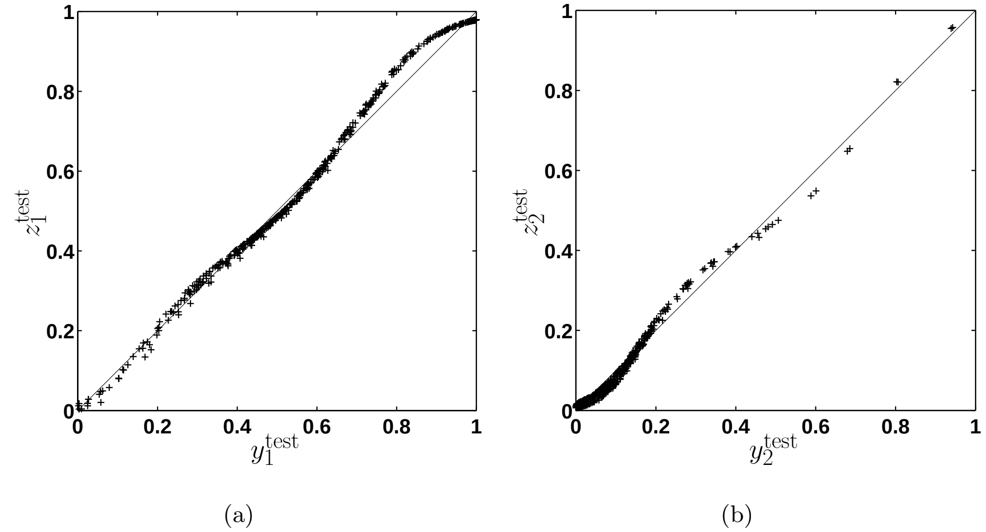


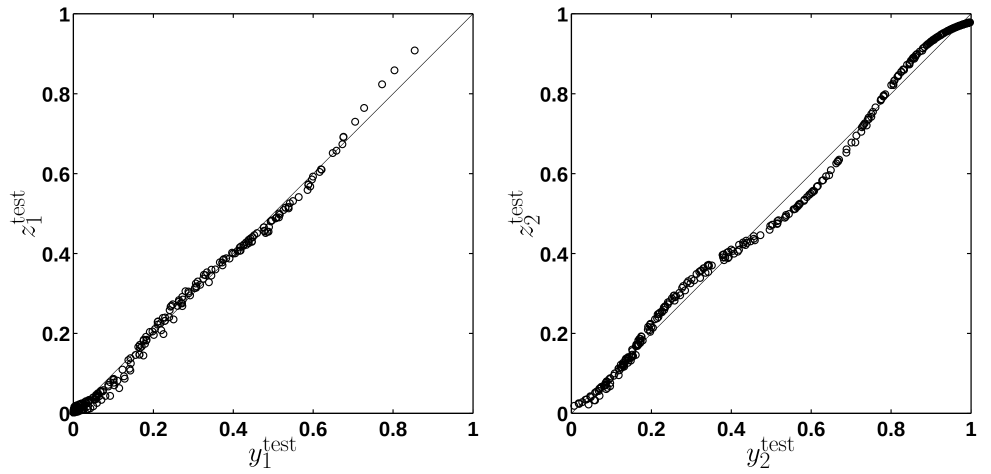
Figure 2. Output of the second two-class BANN as a function of its input for the observations actually drawn from class π_2 in the two-class simulation study.



(a)

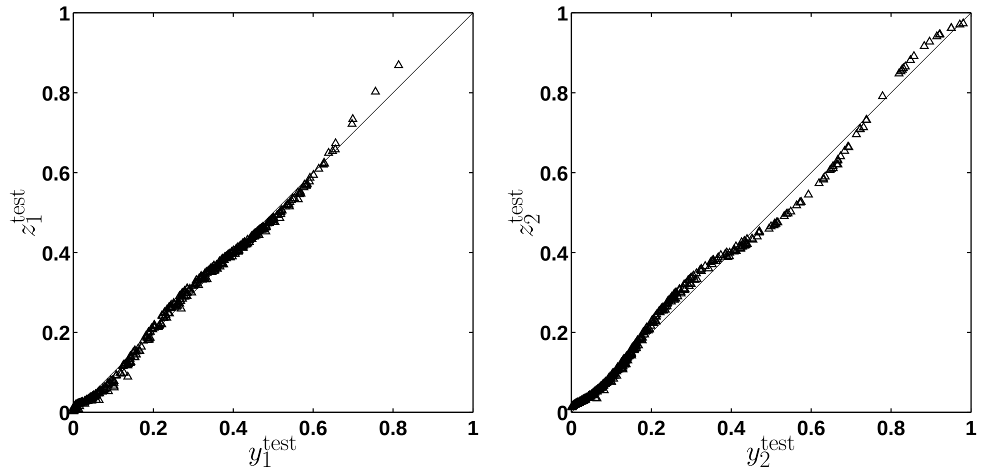
(b)

Figure 3. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_1 in the three-class simulation study.



(a) (b)

Figure 4. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_2 in the three-class simulation study.



(a) (b)

Figure 5. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_3 in the three-class simulation study.

We currently lack a fully general method for three-class classification or for practically evaluating the performance of a three-class classifier. As a first step toward such a classification method, we are investigating the use of BANNs to estimate three-class ideal observer decision variables for such a task. Since, in a practical situation, we will not have access to the underlying probability distributions from which the observational data are drawn, we must rely on circumstantial evidence in support of our claim that a three-class BANN can adequately estimate decision variables directly related to ideal observer decision variables.

Previously, we presented work relating the output of a three-class BANN to the outputs of two-class BANNs trained for various "simplified" cases in which the three-class classification task was reduced to a two-class classification task, and showed that the relationships found were consistent with the relationship between three- and two-class ideal observers for the same tasks.[12] In the present work, we showed that the output of two- and three-class BANNs was consistent, to within experimental error, with the theoretical relationship developed for actual *a posteriori* class membership probabilities. This is of limited practical use in the complete development of a three-class classifier, mainly because the three-class ideal observer decision rule is considerably more complicated than its two-class counterpart (a simple threshold on a single decision variable). It does, however, bolster our confidence in the choice of the BANN as an appropriate tool for estimating the decision variables which would eventually be incorporated in such a classifier.

ACKNOWLEDGMENTS

This work was supported by grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D. C. Edwards, principal investigator). Charles E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

REFERENCES

1. U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.* **2**, pp. 1–9, 1995.
2. F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Med. Phys.* **18**, pp. 955–963, 1991.
3. F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.* **28**, pp. 473–481, 1993.
4. F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.* **21**, pp. 445–452, 1994.
5. M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging.* Ph.D. thesis, The University of Chicago, Chicago, IL, 2000.
6. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.* **5**, pp. 155–168, 1998.
7. Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.* **44**, pp. 2579–2595, 1999.
8. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness," *Acad. Radiol.* **7**, pp. 1077–1084, 2000.
9. Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.* **20**, pp. 1285–1292, 2001.
10. Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis — Observer study with independent database of mammograms," *Radiology* **224**, pp. 560–568, 2002.

11. Z. Huo and M. L. Giger, “Effect of case mix on feature selection in the computerized classification of mammographic lesions,” in Proc. SPIE Vol. 4684 *Medical Imaging 2002: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 762–767, (SPIE, Bellingham, WA), 2002.
12. D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, “Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions,” *Med. Phys.* **31**, pp. 81–90, 2004.
13. D. C. Edwards, M. A. Kupinski, R. H. Nagel, R. M. Nishikawa, and J. Papaioannou, “Using a Bayesian neural network to optimally eliminate false-positive microcalcification detections in a CAD scheme,” in *IWDM 2000: 5th International Workshop on Digital Mammography*, M. J. Yaffe, ed., *Proceedings of the Workshop*, pp. 168–173, Medical Physics Publishing, (Madison, WI), 2001.
14. D. C. Edwards, J. Papaioannou, Y. Jiang, M. A. Kupinski, and R. M. Nishikawa, “Eliminating false-positive microcalcification clusters in a mammography CAD scheme using a Bayesian neural network,” in Proc. SPIE Vol. 4322 *Medical Imaging 2001: Image Processing*, Milan Sonka and Kenneth Hanson, eds., pp. 1954–1960, (SPIE, Bellingham, WA), 2001.
15. D. C. Edwards, C. E. Metz, and R. M. Nishikawa, “Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network,” in Proc. SPIE Vol. 4686 *Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, Dev P. Chakraborty and Elizabeth A. Krupinski, eds., pp. 1–12, (SPIE, Bellingham, WA), 2002.
16. M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, “Ideal observer approximation using Bayesian classification neural networks,” *IEEE Trans. Med. Imag.* **20**, pp. 886–899, 2001.
17. D. J. S. MacKay, *Bayesian Methods for Adaptive Models.* Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.
18. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
19. D. C. Edwards, C. E. Metz, and M. A. Kupinski, “Ideal observers and optimal ROC hypersurfaces in N-class classification,” *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.
20. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., New York, 1991.

E Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule

Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule

Darrin C. Edwards* and Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

We analyzed a variety of recently proposed decision rules for three-class classification from the point of view of ideal observer decision theory. We considered three-class decision rules which have been proposed recently: one by Scurfield, one by Chan *et al.*, and one by Mossman. Scurfield's decision rule can be shown to be a special case of the three-class ideal observer decision rule in two different situations: when the pair of decision variables is the pair of likelihood ratios used by the ideal observer, and when the pair of decision variables is the pair of logarithms of the likelihood ratios. Chan *et al.* start with an ideal observer model, where two of the decision lines used by the ideal observer overlap, and the third line becomes undefined. Finally, we showed that the Mossman decision rule (in which a single decision line separates one class from the other two, while a second line separates those two classes) cannot be a special case of the ideal observer decision rule. Despite the considerable difficulties presented by the three-class classification task compared with two-class classification, we found that the three-class ideal observer provides a useful framework for analyzing a wide variety of three-class decision strategies.

Keywords: ROC analysis, three-class classification, ideal observer decision rules

1. INTRODUCTION

We are attempting to develop a fully automated mass lesion classification scheme for computer-aided diagnosis (CAD) in mammography. This scheme will combine two schemes developed at the University of Chicago: one for automatically detecting mass lesions in mammograms,[1–5] and one for classifying known lesions as malignant or benign.[6–10] Combining these two types of CAD scheme is inherently difficult, because the output of the detection scheme will necessarily include false-positive (FP) computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our approach has been to treat this problem explicitly as a three-class classification task. That is, the outputs of the detection scheme should be classified as malignant lesions, benign lesions, and non-lesions (FP computer detections), and the classifier to be estimated is the ideal observer decision rule for this task. Such an approach presents considerable difficulties of its own. On the one hand, decision rules, in particular ideal observer decision rules, increase rapidly in complexity with the number of classes involved. On the other hand, a fully general performance evaluation method, such as a three-class extension of receiver operating characteristic (ROC) analysis, has yet to be developed.

The explicit form of the ideal observer in a three-class classification task has been known for some time.[11] For the reasons just stated, however, a practical method for estimating and evaluating observer performance based on an ideal observer model has proven elusive, despite the success of the two-class binormal ideal observer model.[12] Nevertheless, pragmatic observer decision rule models for three-class classification tasks have been proposed relatively recently by several groups of researchers. In some cases, these models are motivated more by considerations of tractability than of complete generality. This is of course understandable given the inherent difficulties of three-class classification; however, we thought it might be of interest to analyze a number of recently proposed three-class decision rule models within an ideal observer decision rule framework.

*Correspondence: E-mail: d-edwards@uchicago.edu; Telephone: 773 834 5094; Fax: 773 702 0371

In the next section, we review the three-class ideal observer decision rule. In the following three sections, we review recently proposed three-class decision rule models: one by Scurfield,[13] one by Chan *et al.*,[14] and one by Mossman.[15] In each case, the given decision rule is analyzed in terms of the ideal observer decision rule; where necessary or expedient, assumptions are made about the observer's decision variables in order to facilitate this analysis. We emphasize that we do not attempt a review of the experimental methods in the works discussed; we are specifically interested only in the form of the decision rule which serves as the starting point for each work. The results of our analyses are briefly summarized in Sec. 6.

2. THE THREE-CLASS IDEAL OBSERVER

It can be shown[11, 16] that an N-class ideal observer makes decisions regarding statistically variable observations $\vec{\mathbf{x}}$ by partitioning a likelihood ratio decision variable space, where the boundaries of the partitions are given by hyperplanes:

$$\begin{aligned} &\text{decide} \quad d = \pi_i \quad \text{iff} \\ \sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k) \mathrm{LR}_k \quad &\geq \quad (U_{j|N} - U_{i|N}) P(\mathbf{t} = \pi_N) \qquad \{j < i\} \qquad (1) \\ &\text{and} \\ \sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k) \mathrm{LR}_k \quad &> \quad (U_{j|N} - U_{i|N}) P(\mathbf{t} = \pi_N) \qquad \{j > i\}. \qquad (2) \end{aligned}$$

Here $U_{i|j}$ is the utility of deciding an observation is from class π_i given that it is actually from class π_j, and the $N-1$ likelihood ratios are defined as

$$\mathrm{LR}_i \equiv \frac{p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_i)}{p_{\vec{x}}(\vec{x}|\mathbf{t} = \pi_N)} \qquad (3)$$

for $i < N$. We also define the actual class (the "truth") to which an observation belongs as $\mathbf{t}$, and the class to which it is assigned (the "decision") as $\mathbf{d}$, where $\mathbf{t}$ and $\mathbf{d}$ can take on any of the values $\pi_1, \ldots, \pi_i, \ldots, \pi_N$, the labels of the various classes. (We use boldface type to denote statistically variable quantities.)

The partitioning of the decision variable space is determined by the parameters

$$\gamma_{ijk} \equiv (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k), \qquad (4)$$

with i, j, and k varying from 1 to N, and $j \neq i$. Note that these parameters are not independent, however, because

$$\gamma_{ijk} = \gamma_{kjk} - \gamma_{kik}. \qquad (5)$$

We can impose the reasonable condition that the utility for correctly classifying an observation from a given class should be greater than any utility for incorrectly classifying an observation from the same class, *i.e.*, $U_{i|i} > U_{j|i} \quad \{i \neq j\}$. This gives, for $j \neq i$,

$$\gamma_{iji} > 0, \qquad (6)$$

leaving $N(N-1)$ parameters (the rest are derivable from Eq. 5).

Finally, note that the hyperplanes represented by Eqs. 1 and 2 are unchanged if we multiply all of these equations by a single scalar, such as $1/(\sum_{i \neq j} \gamma_{iji})$. This leaves us with $N^2 - N - 1$ degrees of freedom, as expected.

The behavior of a three-class ideal observer is completely determined by the three decision boundary lines

$$\begin{aligned} \gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 &= \gamma_{313} - \gamma_{323} \qquad (7) \\ \gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 &= \gamma_{313} \qquad (8) \\ (\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 &= \gamma_{323}, \qquad (9) \end{aligned}$$

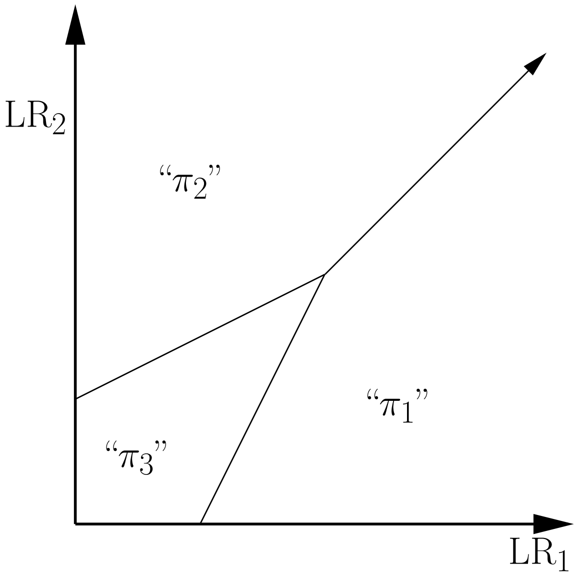


Figure 1. Example three-class ideal observer decision rule, given the values of the decision parameters $\gamma_{121} = \gamma_{212} = 3/14$ and $\gamma_{131} = \gamma_{313} = \gamma_{232} = \gamma_{323} = 1/7$. Note $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\mathbf{t} = \pi_k)$.

which we call, respectively, the "1-*vs.*-2" line, the "1-*vs.*-3" line, and the "2-*vs.*-3" line. Note that if any two of these lines intersect, the third line must also share this intersection point. We also emphasize the simple interpretation, from Eq. 4, of each of the γ_{iji} parameters appearing in these decision boundary line equations as the difference in utilities between a "correct" and one particular "incorrect" decision (scaled by the *a priori* probability of the true class in question); and of each difference in the γ_{iji} parameters as a difference in utilities between two possible "incorrect" decisions (again scaled by the *a priori* probability of the true class in question).

An example ideal observer decision rule for particular values of the utilities $U_{i|j}$, and hence of the parameters γ_{iji}, is shown in Fig. 1. Here we have chosen $\gamma_{121} = \gamma_{212} = 3/14$ and $\gamma_{131} = \gamma_{313} = \gamma_{232} = \gamma_{323} = 1/7$, yielding the decision boundary lines

$$\frac{3}{14}\mathrm{LR}_1 - \frac{3}{14}\mathrm{LR}_2 = 0 \quad \{\text{"1-}\textit{vs.}\text{-2"}\} \tag{10}$$

$$\frac{1}{7}\mathrm{LR}_1 - \frac{1}{14}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{"1-}\textit{vs.}\text{-3"}\} \tag{11}$$

$$-\frac{1}{14}\mathrm{LR}_1 + \frac{1}{7}\mathrm{LR}_2 = \frac{1}{7} \quad \{\text{"2-}\textit{vs.}\text{-3"}\}. \tag{12}$$

These simplify to the equations $\mathrm{LR}_2 = \mathrm{LR}_1$, $\mathrm{LR}_2 = 2\mathrm{LR}_1 - 2$, and $\mathrm{LR}_2 = \mathrm{LR}_1/2 + 1$, respectively.

3. THE SCURFIELD DECISION RULE

Scurfield investigated a decision rule applied to two-dimensional statistically variable data ($\vec{\mathbf{y}} \equiv (\mathbf{y}_1, \mathbf{y}_2)$) drawn from three classes.[13] The application domain was human observer performance modeling for acoustical psychophysics experiments. (In prior work, Scurfield investigated a decision rule for three-class classification of univariate data.[17] We will not review that prior work here, because at present we are interested in relating given observer models to the three-class ideal observer model for multivariate observational data, which yield two-dimensional decision variable data by Eq. 3.) In Scurfield's work, no assumptions are made about the decision variables $\mathbf{y}_1$ and $\mathbf{y}_2$; in particular, these decision variables are not assumed to be related in any way to an ideal observer model. This is entirely appropriate given the nature of the problem domain Scurfield investigated — *i.e.*, human observer performance modeling. It can readily be shown, however, that if one chooses to make such assumptions, special cases of the Scurfield model are in fact special cases of an ideal observer decision rule.

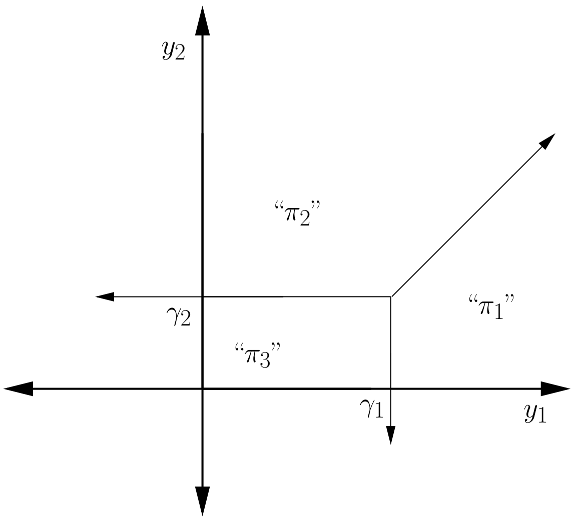


Figure 2. Decision rule investigated by Scurfield, for the decision parameters γ_1 and γ_2.

The Scurfield decision rule is dependent on two decision parameters, which we will call γ_1 and γ_2. The decision rule can be written as

$$\begin{aligned} &\text{decide} \quad d=\pi_1 \quad \text{iff} \quad y_1-y_2 \geq \gamma_1-\gamma_2 \quad \text{and} \quad y_1 \geq \gamma_1; && (13)\\ &\text{decide} \quad d=\pi_2 \quad \text{iff} \quad y_1-y_2 < \gamma_1-\gamma_2 \quad \text{and} \quad y_2 \geq \gamma_2; && (14)\\ &\text{decide} \quad d=\pi_3 \quad \text{iff} \quad y_1 < \gamma_1 \quad \text{and} \quad y_2 < \gamma_2. && (15)\end{aligned}$$

This decision rule is illustrated in Fig. 2.

From these relations, one can define the decision boundary lines

$$\begin{aligned} y_1-y_2 &= \gamma_1-\gamma_2 \quad \{\text{“1-}vs.\text{-2”}\} && (16)\\ y_1 &= \gamma_1 \quad \{\text{“1-}vs.\text{-3”}\} && (17)\\ y_2 &= \gamma_2 \quad \{\text{“2-}vs.\text{-3”}\}. && (18)\end{aligned}$$

Note the similarity in form between these equations and Eqs. 7-9. If we choose $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$ for some set of observational data $\vec{\mathbf{x}}$, we have a special case of Eqs. 7-9, which is illustrated in Fig. 3.

A second correspondence between Scurfield's decision rule and the ideal observer decision rule can be obtained by taking $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$; note that a line of the form $\log(\mathrm{LR}_2) = \log(\mathrm{LR}_1) + \alpha$ corresponds to a line of the form $\mathrm{LR}_2 = \beta \mathrm{LR}_1$ for appropriate constants α and β. By inspection, this is again a special case of Eqs. 7-9, which is illustrated in Fig. 4.

Scurfield points out[13] that the observer which maximizes P_C, the "percent correct" or probability of a correct response, is a special case of the ideal observer (*i.e.*, a single operating point achievable by the ideal observer for the given task). This observer follows the Scurfield decision rule model with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$, and decision parameters given by $e^{\gamma_1} = P(\pi_3)/P(\pi_1)$ and $e^{\gamma_2} = P(\pi_3)/P(\pi_2)$. It is interesting to note that the Scurfield decision rule model can in fact be used to describe ideal observer performance for an even wider class of operating points, as shown in this section.

4. THE CHAN DECISION RULE

Chan *et al.* are investigating three-class classifiers for computer-aided diagnosis.[14] Their work is motivated by reasoning similar in principle to that which we independently arrived at when we began to consider this problem. In particular, they consider a clinical situation in which observations must be classified as malignant, benign,

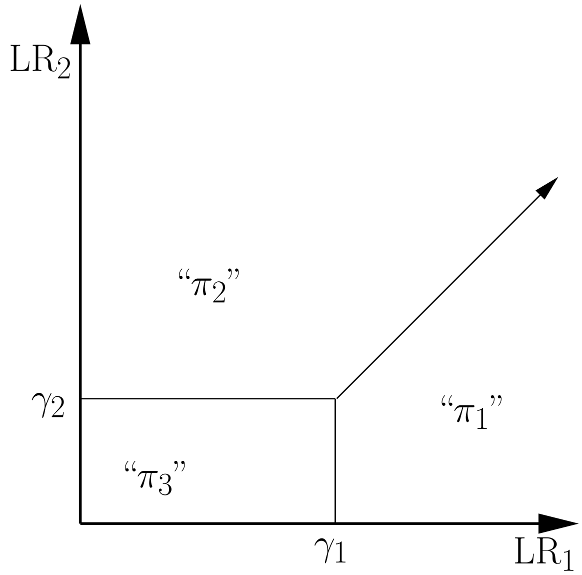


Figure 3. A special case of the ideal observer decision rule, which is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$.

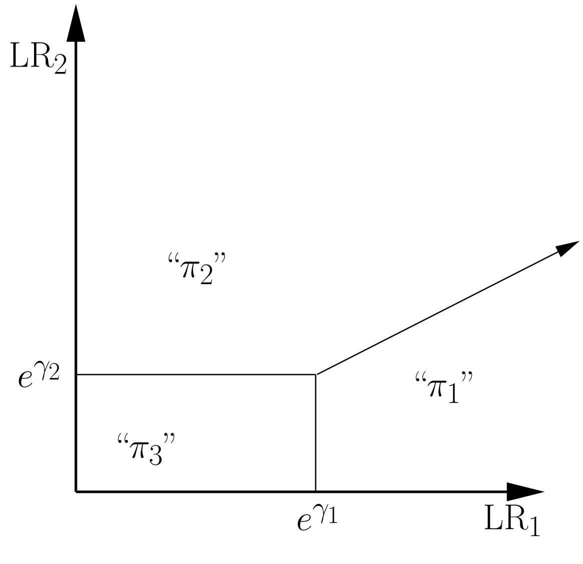


Figure 4. A special case of the ideal observer decision rule which is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$.

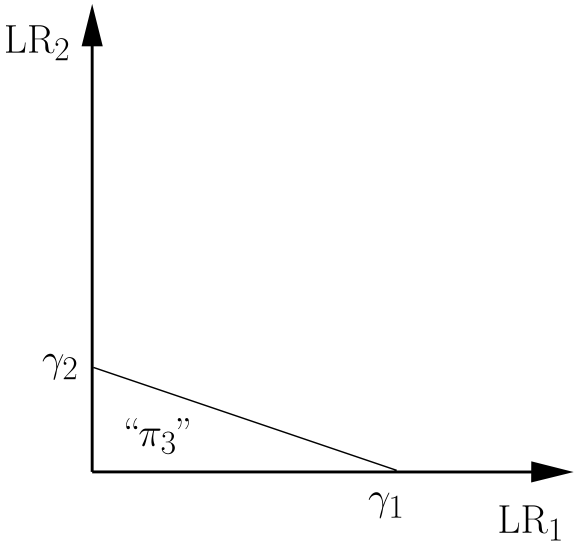


Figure 5. The decision rule investigated by Chan *et al.*, which as they state is a special case of the ideal observer decision rule. Observations in the unlabelled region are decided "not π_3", *i.e.*, either "π_1" or "π_2".

or normal. Because the goal of their work is to optimize the performance of a system to aid a radiologist or clinician, rather than to measure the psychophysical performance of an existing observer, they choose to start explicitly from an ideal observer model in constructing their decision rule.

In order to reduce the complexity of the ideal observer decision rule to manageable proportions, Chan *et al.* impose restrictions on the utilities used by their observer. In their formulation, the class we are labelling π_1 is the benign class; π_2, the normal class; and the malignant class is π_3. They further assume that the possible values of any utility $U_{i|j}$ are restricted to the interval $[0, 1]$. They then set $U_{1|1} = U_{2|2} = U_{3|3} = 1$ (*i.e.*, correctly identifying any case has maximal utility). Furthermore, they require $U_{2|1} = U_{1|2} = 1$ and $U_{1|3} = U_{2|3} = 0$ (*i.e.*, misidentifying a benign case as normal, or vice versa, has no significant cost reducing the utility of such a decision from the maximum, but misclassifying an actually malignant case as benign or normal has the minimum possible utility). Finally, $U_{3|1}$, and $U_{3|2}$ are assumed to have arbitrary values on the open interval $(0, 1)$ (*i.e.*, misclassifying an actually non-malignant case as malignant will have some cost reducing the utility of such a decision from the maximum, but such a misclassification is in some sense "better" than missing an actual malignancy). It is important to note that these assumptions are arguably relevant to a reasonable model of a clinical situation, and are thus of interest beyond their superficial advantage in reducing the degrees of freedom involved in the observer's decision rule. We will, however, only consider the latter issue in the remainder of this section.

Substituting the values of the utilities given above into Eq. 4, we obtain decision boundary lines of the form

$$0\,\mathrm{LR}_1 + 0\,\mathrm{LR}_2 = 0 \qquad \{\text{“1-}\textit{vs.}\text{-2”}\} \tag{19}$$

$$\frac{(1-U_{3|1})P(\mathbf{t}=\pi_1)}{\alpha}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\mathbf{t}=\pi_2)}{\alpha}\mathrm{LR}_2 = \frac{P(\mathbf{t}=\pi_3)}{\alpha} \qquad \{\text{“1-}\textit{vs.}\text{-3”}\} \tag{20}$$

$$\frac{(1-U_{3|1})P(\mathbf{t}=\pi_1)}{\alpha}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\mathbf{t}=\pi_2)}{\alpha}\mathrm{LR}_2 = \frac{P(\mathbf{t}=\pi_3)}{\alpha} \qquad \{\text{“2-}\textit{vs.}\text{-3”}\} \tag{21}$$

where $\alpha \equiv 1 + P(\mathbf{t} = \pi_3) - U_{3|1}P(\mathbf{t} = \pi_1) - U_{3|2}P(\mathbf{t} = \pi_2)$. Note that, as Chan *et al.* point out, the "1-*vs.*-2" line is in fact undefined for this choice of utilities, while the "1-*vs.*-3" and "2-*vs.*-3" lines are identical. This is a general consequence of Eqs. 7-9; if any two of these equations yield identical lines, the third line must be either identical to them or undefined. The decision rule considered by Chan *et al.* is illustrated in Fig. 5.

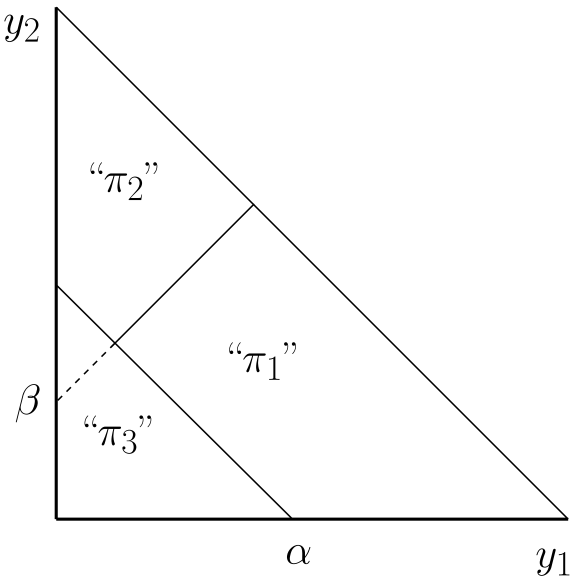


Figure 6. Decision rule investigated by Mossman, for the decision parameters α and β, shown in the *a posteriori* class probability space.

5. THE MOSSMAN DECISION RULE

Mossman investigates a decision rule applied to a set of three decision variables $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$, subject to the constraint

$$\mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 = 1, \tag{22}$$

as well as $0 \leq \mathbf{y}_i \leq 1 \;\; \{1 \leq i \leq 3\}$. This is consistent with the constraint on the *a posteriori* class probabilities, $P(\pi_1|\vec{\mathbf{x}}) + P(\pi_2|\vec{\mathbf{x}}) + P(\pi_3|\vec{\mathbf{x}}) = 1$; these quantities are known to be directly related to the likelihood ratio ideal observer decision variables.[18, 19] (In this section we will write $P(\pi_i|\vec{x})$ instead of $P(\mathbf{t} = \pi_i|\vec{x})$ for simplicity.) Mossman does not explicitly require, however, that the decision variables in Eq. 22 be the *a posteriori* class probabilities (*e.g.*, they may be noisy estimates of these quantities).

The decision rule considered by Mossman, which depends on two decision parameters α and β, is

$$\begin{array}{llllll} \text{decide} & d = \pi_1 & \text{iff} & y_2 - y_1 \leq \beta & \text{and} & y_3 \leq \alpha; \end{array} \tag{23}$$
$$\begin{array}{llllll} \text{decide} & d = \pi_2 & \text{iff} & y_2 - y_1 > \beta & \text{and} & y_3 \leq \alpha; \end{array} \tag{24}$$
$$\begin{array}{llll} \text{decide} & d = \pi_3 & \text{iff} & y_3 > \alpha. \end{array} \tag{25}$$

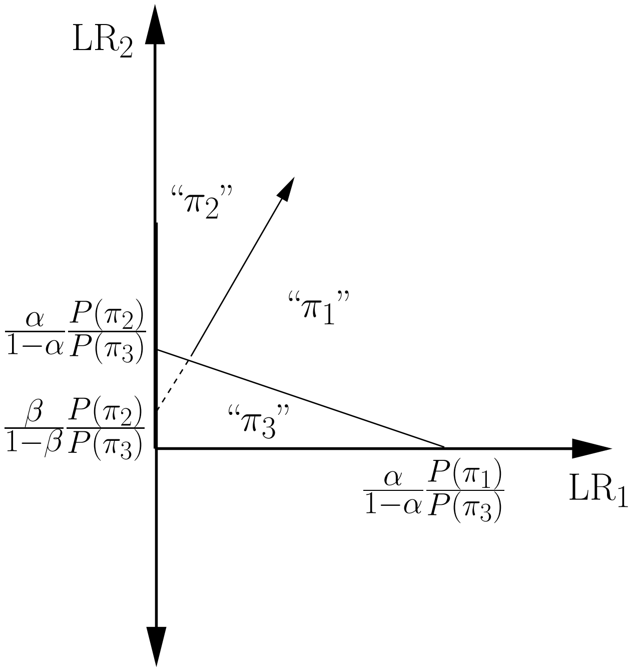
where $0 \leq \alpha \leq 1$ and $-1 \leq \beta \leq 1$. From these relations, and given the relation $y_3 = 1 - y_1 - y_2$ from Eq. 22, one can define the decision boundary lines

$$y_1 - y_2 = -\beta \quad \{\text{“1-}vs.\text{-2”}\} \tag{26}$$
$$y_1 + y_2 = 1 - \alpha \quad \{\text{“1-}vs.\text{-3”}\} \tag{27}$$
$$y_1 + y_2 = 1 - \alpha \quad \{\text{“2-}vs.\text{-3”}\}. \tag{28}$$

This decision rule is illustrated in Fig. 6. Note that, similar to the Chan *et al.* decision rule, the "1-*vs.*-3" and "2-*vs.*-3" decision boundary lines are identical.

We now consider a special case of the Mossman decision rule in which $\mathbf{y}_1 = P(\pi_1|\vec{\mathbf{x}})$, $\mathbf{y}_2 = P(\pi_2|\vec{\mathbf{x}})$, and $\mathbf{y}_3 = P(\pi_3|\vec{\mathbf{x}})$ for some observational data vector $\vec{\mathbf{x}}$. This version of the decision rule is illustrated in Fig. 7.

Although the Mossman decision rule appears similar in form to the ideal observer decision rule, recall from Sec. 4 that if two of the decision boundary line equations are identical, the third must yield a line identical to



LR_2
"π_2"
"π_1"
$\frac{\alpha}{1-\alpha}\frac{P(\pi_2)}{P(\pi_3)}$
"π_3"
$\frac{\beta}{1-\beta}\frac{P(\pi_2)}{P(\pi_3)}$
$\frac{\alpha}{1-\alpha}\frac{P(\pi_1)}{P(\pi_3)}$
LR_1

Figure 7. Decision rule investigated by Mossman, for the decision parameters α and β, shown in likelihood ratio space.

the first two or be undefined. Another way to see this is to note that the coefficients of Eq. 9 are differences of the corresponding coefficients of Eqs. 7 and 8. If the coefficients of Eqs. 8 and 9 are identical, it must be the case that the coefficients of Eq. 7 are all zero. For the Mossman decision rule, this would require $1+\beta=0$, $1-\beta=0$, and $\beta=0$ simultaneously, which is clearly impossible. It follows that the decision rule considered by Mossman cannot represent possible ideal observer performance for any choice of the utilities $U_{i|j}$ in Eqs. 1 and 2.

6. DISCUSSION AND CONCLUSIONS

We examined three decision rules proposed recently for three-class classification tasks by different researchers. The basis for our evaluation was ideal observer decision theory, primarily because our own interest in the three-class classification task is its possible application to CAD.

Although this is not the most general approach to three-class classification, the three-class classification task is difficult enough that it is perhaps worth making any attempt to analyze, from a single point of view, the work of the relatively few researchers investigating this problem.

In particular, Scurfield points out[13] that his proposed decision rule is in fact an ideal observer decision rule for a single ideal observer operating point, namely the observer which maximizes the probability of any correct response (or "percent correct" or P_C). We were able to show that, under various assumptions, a larger set of such correspondences between the Scurfield observer and the ideal observer exists.

Chan *et al.* are working on the application of three-class classification to CAD, and thus explicitly take the ideal observer as the starting point in the development of their decision rule.[14] Although this rendered our analysis of that decision rule in terms of ideal observer decision theory largely trivial, it provided an intuitive basis for understanding the results of similar analysis of the Mossman decision rule, namely the conclusion that the latter does not correspond to ideal observer behavior for any possible values of the utilities used by the ideal observer. However, we note that the structure of the Mossman decision rule — a simple sequence of thresholds on single decision variables — may indeed serve as a reasonable model for human observer performance in certain situations, *e.g.*, differential diagnosis.

ACKNOWLEDGMENTS

The authors thank Vit Drga for bringing Brian Scurfield's work to their attention, and thank both Vit Drga and Brian Scurfield for helpful conversations concerning that work. The authors thank Heang-Ping Chan and Berkman Sahiner for helpful conversations concerning their work.

This work was supported in parts by grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D. C. Edwards, principal investigator). Charles E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

REFERENCES

1. U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.* **2**, pp. 1–9, 1995.
2. F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Med. Phys.* **18**, pp. 955–963, 1991.
3. F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.* **28**, pp. 473–481, 1993.
4. F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.* **21**, pp. 445–452, 1994.
5. M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging.* Ph.D. thesis, The University of Chicago, Chicago, IL, 2000.
6. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.* **5**, pp. 155–168, 1998.
7. Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.* **44**, pp. 2579–2595, 1999.
8. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness," *Acad. Radiol.* **7**, pp. 1077–1084, 2000.
9. Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.* **20**, pp. 1285–1292, 2001.
10. Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis — Observer study with independent database of mammograms," *Radiology* **224**, pp. 560–568, 2002.
11. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
12. C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, pp. 1–33, 1999.
13. B. K. Scurfield, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psychol.* **42**, pp. 5–31, 1998.
14. H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 567–578, (SPIE, Bellingham, WA), 2003.
15. D. Mossman, "Three-way ROCs," *Med. Decis. Making* **19**, pp. 78–89, 1999.
16. D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.

17. B. K. Scurfield, “Multiple-event forced-choice tasks in the theory of signal detectability,” *J. Math Psychol.* **40**, pp. 253–269, 1996.
18. M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, “Ideal observer approximation using Bayesian classification neural networks,” *IEEE Trans. Med. Imag.* **20**, pp. 886–899, 2001.
19. D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, “Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions,” *Med. Phys.* **31**, pp. 81–90, 2004.

F Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule

Available online at www.sciencedirect.com

ScienceDirect

Journal of Mathematical Psychology 50 (2006) 478–487

Journal of Mathematical Psychology

www.elsevier.com/locate/jmp

Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule[☆]

Darrin C. Edwards*, Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637, USA

Received 19 May 2005; received in revised form 9 May 2006
Available online 1 August 2006

Abstract

We analyze recently proposed decision rules for three-class classification from the point of view of ideal observer decision theory. We consider three-class decision rules proposed by Scurfield, by Chan et al., and by Mossman. Scurfield's decision rule is shown to be a special case of the three-class ideal observer decision rule in three different situations. Chan et al. start with an ideal observer model and specify its decision-consequence utility structure in a way that causes two of the decision lines used by the ideal observer to overlap and the third line to become undefined. Finally, we show that, for a particular and obvious choice of ideal-observer-related decision variables, the Mossman decision rule cannot be a special case of the ideal observer decision rule. Despite the considerable difficulties presented by the three-class classification task, the three-class ideal observer provides a useful framework for analyzing a variety of three-class decision strategies.

Keywords: ROC analysis; Three-class classification; Ideal observer decision rules

1. Introduction

We are attempting to develop a fully automated mass lesion classification scheme for computer-aided diagnosis (CAD) in mammography. This scheme will combine two schemes developed at the University of Chicago: one for automatically detecting mass lesions in mammograms (Bick et al., 1995; Kupinski, 2000; Yin et al., 1991, Yin, Giger, Vyborny, Doi, & Schmidt, 1993, 1994), and one for classifying known lesions as malignant or benign (Huo, Giger, & Metz, 1999; Huo, Giger, & Vyborny, 2001; Huo, Giger, Vyborny, & Metz, 2002; Huo, Giger, Vyborny, Wolverton, & Metz, 2000; Huo et al., 1998). Combining these two types of CAD scheme is inherently difficult, because the output of the detection scheme will necessarily include false-positive (FP) computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our approach has been to treat this problem explicitly as a three-class classification task. That is, the outputs of the detection scheme should be classified as malignant lesions, benign lesions, and non-lesions (FP computer detections), and the classifier to be estimated is the ideal observer decision rule for this task. Such an approach presents considerable difficulties of its own. On the one hand, decision rules, in particular ideal observer decision rules, increase rapidly in complexity with the number of classes involved. On the other hand, a fully general performance evaluation method, such as a three-class extension of receiver operating characteristic (ROC) analysis, has yet to be developed. It should be mentioned that the simple model we have just described corresponds in the two-class classification task to ROC analysis performed "per

[☆] This work was supported by Grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D.C. Edwards, principal investigator). C.E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

*Corresponding author. Fax: +1 773 702 0371.
E-mail address: d-edwards@uchicago.edu (D.C. Edwards).

0022-2496/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2006.05.004

detection;" that is, each "case" being classified corresponds to a small region of interest (ROI) in the image containing a single computer detection. Other formulations, such as ROC analysis "per image," ROC analysis "per patient" (for a set of images, such as the four mammographic views obtained in a typical screening setting), or free-response ROC (FROC) (Bunch, Hamilton, Sanderson, & Simmons, 1978; Chakraborty, 1989, 2002) analysis, are also possible, but their extension to tasks with three or more classes is beyond the scope of the present work.

The explicit form of the decision rule used by the ideal observer in a three-class classification task has been known for some time (Van Trees, 1968). For the reasons just stated, however, a practical and general method for estimating and evaluating observer performance has proven elusive. In particular, Scurfield (1996) defined the two-class information-based performance metric $D_{1:2} \equiv \log 2 - \text{AUC} \log \text{AUC} - (1 - \text{AUC}) \log(1 - \text{AUC})$ (where AUC is the area under the two-class ROC curve), and extended it to the three-class case for two different decision rules (Scurfield, 1996, 1998). Srinivasan (1999) investigated the optimality of discrete, multi-class ROC operating points, but not continuous ROC hypersurfaces, under a cost function equivalent to the Bayes risk. Mossman (1999) evaluated the performance of a three-class classifier with a surface formed from the three correct classification probabilities. Hand and Till (2001) proposed the average of the areas under all $N(N-1)/2$ between-class ROC curves as a performance metric in an N-class classification task. Obuchowski et al. (2001) elicited readers' estimates of the set of probabilities of each observation belonging to N classes, and then used conventional (two-class) ROC analysis to evaluate each of the $N(N-1)/2$ differences of these estimates for its ability to distinguish between the relevant pair of classes. Ferri, Hernández-Orallo, and Salido (2003) proposed a variety of algorithms for calculating the hypervolume under the convex hull obtained from a set of discrete ROC operating points; a modified version of the Hand and Till metric averaging the N areas under the ROC surfaces that measure the observer's ability to distinguish a given class from the remaining $N-1$; and a graphical "cobweb" representation of the observer's misclassification probabilities. Lachiche and Flach (2003) proposed iterative algorithms for finding the optimal among a discrete set of multi-class ROC operating points based on either percent correct or Bayes risk. Nakas and Yiannoutsos (2004) considered an observer using a decision rule similar to that of Scurfield (1996), and evaluated its performance statistically by extending methods proposed by Dreiseitl, Ohno-Machado, and Binder (2000). Patel and Markey (2005) applied a variety of proposed evaluation metrics, including the Hand and Till metric, the modified Hand and Till metric of Ferri, the "cobweb" graphical measure of Ferri, and the Mossman ROC surface, to radiologist assessment data of mammographic images from patients who subsequently underwent biopsy.

The works cited above demonstrate the difficulty in developing a fully general performance metric for classification tasks with more than two classes. Lacking such a performance metric in turn makes the development of observer decision rules for such tasks difficult, because they can at present be evaluated and compared only from a theoretical rather than an empirical perspective. Nevertheless, observer decision rule models for three-class classification tasks have been proposed relatively recently by several groups of researchers. In some cases, these models are motivated more by considerations of tractability than of complete generality. This is of course understandable given the inherent difficulties of three-class classification; however, we thought it might be of interest to analyze a number of recently proposed three-class decision rule models within an ideal observer decision rule framework.

In the next section, we review the three-class ideal observer decision rule. In the following three sections, we review recently proposed three-class decision rule models: one by Scurfield (1998), one by Chan, Sahiner, Hadjiiski, Petrick, and Zhou (2003), and one by Mossman (1999). In each case, the given decision rule is analyzed in terms of the ideal observer decision rule; where necessary or expedient, assumptions are made about the observer's decision variables in order to facilitate this analysis. We emphasize that we do not attempt a review of the experimental methods or detailed analysis of proposed performance evaluation metrics in the works discussed; we are here interested only in the form of the decision rule which serves as the starting point for each work, and superficially in the proposed evaluation metrics inasmuch as they are related to those decision rules. (Because of the lack of a fully general performance metric, or figure of merit, for the three-class classification task and, in particular, apparent inconsistencies which are obtained from a straightforward generalization of the area under the ROC curve (Edwards, Metz, & Nishikawa, 2005) we do not attempt any validation or quantitative comparison of the proposed performance metrics.) The results of our analyses are briefly summarized in Section 6.

2. The three-class ideal observer

It can be shown (Edwards, Metz, & Kupinski, 2004b; Van Trees, 1968) that an N-class ideal observer makes decisions regarding statistically variable observations $\vec{\mathbf{x}}$ by partitioning a likelihood ratio decision variable space, where the boundaries of the partitions are given by hyperplanes

decide $d = \pi_i$ iff

$$\sum_{k=1}^{N-1} (U_{i|k} - U_{j|k}) P(\mathbf{t} = \pi_k) \mathrm{LR}_k \geqslant (U_{j|N} - U_{i|N}) P(\mathbf{t} = \pi_N) \;\; \{j < i\} \qquad (1)$$

and

$$\sum_{k=1}^{N-1}(U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k)\mathrm{LR}_k > (U_{j|N}-U_{i|N})P(\mathbf{t}=\pi_N)\ \{j>i\}. \tag{2}$$

Here $U_{i|j}$ is the utility of deciding an observation is from class π_i given that it is actually from class π_j, and the $N-1$ likelihood ratios are defined as

$$\mathrm{LR}_k \equiv \frac{p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_k)}{p_{\vec{x}}(\vec{x}|\mathbf{t}=\pi_N)} \tag{3}$$

for $k<N$. We also define the actual class (the "truth") to which an observation belongs as **t**, and the class to which it is assigned (the "decision") as **d**, where **t** and **d** can take on any of the values $\pi_1,\ldots,\pi_i,\ldots,\pi_N$, the labels of the various classes. (We use boldface type to denote statistically variable quantities.) For simplicity, we will usually write π_k to denote the event $\mathbf{t}=\pi_k$, as in the a priori probability $P(\pi_k)$.

The partitioning of the decision variable space is determined by the parameters

$$\gamma_{ijk} \equiv (U_{i|k}-U_{j|k})P(\pi_k), \tag{4}$$

with i, j, and k varying from 1 to N, and $j\neq i$. Note that these parameters are not independent, however, because

$$\gamma_{ijk} = \gamma_{kjk} - \gamma_{kik}. \tag{5}$$

We can impose the reasonable condition that the utility for correctly classifying an observation from a given class should be greater than any utility for incorrectly classifying an observation from the same class, i.e., $U_{i|i} > U_{j|i}\quad \{i\neq j\}$. This gives, for $j\neq i$,

$$\gamma_{iji} > 0, \tag{6}$$

leaving $N(N-1)$ parameters (the rest are derivable from (5)).

Finally, note that the hyperplanes represented by (1) and (2) are unchanged if we multiply all of these relations by a single scalar, such as $1/(\sum_{i\neq j}\gamma_{iji})$. This leaves us with N^2-N-1 degrees of freedom, as expected, and effectively imposes the condition

$$\sum_{i\neq j}\gamma_{iji} = 1. \tag{7}$$

The behavior of a three-class ideal observer is completely determined by the three decision boundary lines

$$\gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 = \gamma_{313} - \gamma_{323}, \tag{8}$$

$$\gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 = \gamma_{313}, \tag{9}$$

$$(\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 = \gamma_{323}, \tag{10}$$

which we call, respectively, the "1-*vs.*-2" line, the "1-*vs.*-3" line, and the "2-*vs.*-3" line. Note that if any two of these lines intersect, the third line must also share this intersection point. We also emphasize the simple interpretation, from (4), of each of the γ_{iji} parameters appearing in these decision boundary line equations as the difference in utilities between a "correct" and one particular "incorrect" decision (scaled by the a priori probability of the true class in question); and of each difference in the γ_{iji} parameters as a difference in utilities between two possible "incorrect" decisions (again scaled by the a priori probability of the true class in question).

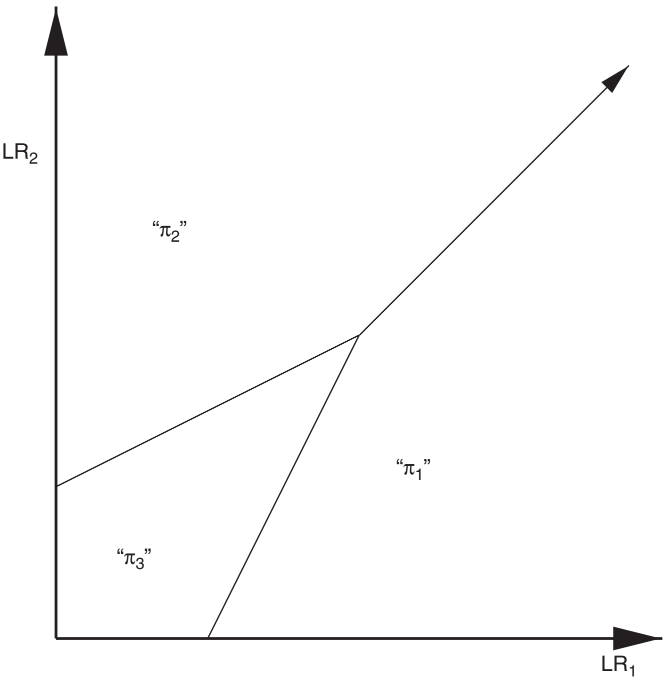


Fig. 1. Example three-class ideal observer decision rule, given the values of the decision parameters $\gamma_{121}=\gamma_{212}=\frac{3}{14}$ and $\gamma_{131}=\gamma_{313}=\gamma_{232}=\gamma_{323}=\frac{1}{7}$. Note that $\gamma_{iji}\equiv(U_{i|i}-U_{j|i})P(\mathbf{t}=\pi_i)$.

An example ideal observer decision rule for particular values of the utilities $U_{i|j}$, and hence of the parameters γ_{iji}, is shown in Fig. 1. Here we have chosen $\gamma_{121}=\gamma_{212}=\frac{3}{14}$ and $\gamma_{131}=\gamma_{313}=\gamma_{232}=\gamma_{323}=\frac{1}{7}$, yielding the decision boundary lines

$$\tfrac{3}{14}\mathrm{LR}_1 - \tfrac{3}{14}\mathrm{LR}_2 = 0 \quad \{\text{" 1-}vs.\text{-2"}\}\,, \tag{11}$$

$$\tfrac{1}{7}\mathrm{LR}_1 - \tfrac{1}{14}\mathrm{LR}_2 = \tfrac{1}{7} \quad \{\text{" 1-}vs.\text{-3"}\}\,, \tag{12}$$

$$-\tfrac{1}{14}\mathrm{LR}_1 + \tfrac{1}{7}\mathrm{LR}_2 = \tfrac{1}{7} \quad \{\text{" 2-}vs.\text{-3"}\}\,. \tag{13}$$

These simplify to the equations $\mathrm{LR}_2=\mathrm{LR}_1$, $\mathrm{LR}_2=2\mathrm{LR}_1-2$, and $\mathrm{LR}_2=\mathrm{LR}_1/2+1$, respectively.

3. The Scurfield decision rule

Scurfield investigated a decision rule applied to two-dimensional statistically variable data ($\vec{\mathbf{y}}\equiv(\mathbf{y}_1,\mathbf{y}_2)$) drawn from three classes (Scurfield, 1998). The application domain was human observer performance modeling for acoustical psychophysics experiments. (In prior work, Scurfield investigated a decision rule for three-class classification of univariate data (Scurfield, 1996). We will not review that prior work here, because at present we are interested in relating given observer models to the general three-class ideal observer model for multivariate observational data, which—except in degenerate cases—will yield two-dimensional decision variable data by (3).) In Scurfield's work, no assumptions are made about the decision variables $\mathbf{y}_1$ and $\mathbf{y}_2$; in particular, these decision variables are not assumed to be related in any way to an ideal observer model. This is entirely appropriate given the nature of the problem domain Scurfield investigated—i.e., human observer performance modeling. It can readily be shown, however, that if one chooses to make such assumptions, special cases of the Scurfield model are in fact special cases of an ideal observer decision rule.

The Scurfield decision rule is dependent on two decision parameters, which we will call γ_1 and γ_2. The decision rule can be written as

$$\text{decide} \quad d = \pi_1 \quad \text{iff } y_1 - y_2 \geqslant \gamma_1 - \gamma_2 \quad \text{and} \quad y_1 \geqslant \gamma_1, \tag{14}$$

$$\text{decide} \quad d = \pi_2 \quad \text{iff } y_1 - y_2 < \gamma_1 - \gamma_2 \quad \text{and} \quad y_2 \geqslant \gamma_2, \tag{15}$$

$$\text{decide} \quad d = \pi_3 \quad \text{iff } y_1 < \gamma_1 \quad \text{and} \quad y_2 < \gamma_2. \tag{16}$$

This decision rule is illustrated in Fig. 2.

From these relations, one can define the decision boundary lines

$$y_1 - y_2 = \gamma_1 - \gamma_2 \quad \{\text{`` 1-}vs.\text{-2''}\}, \tag{17}$$

$$y_1 = \gamma_1 \quad \{\text{`` 1-}vs.\text{-3''}\}, \tag{18}$$

$$y_2 = \gamma_2 \quad \{\text{`` 2-}vs.\text{-3''}\}. \tag{19}$$

If we choose $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$ for some set of observational data $\vec{\mathbf{x}}$, we have

$$\frac{1}{\gamma_0}\mathrm{LR}_1 - \frac{1}{\gamma_0}\mathrm{LR}_2 = \frac{\gamma_1 - \gamma_2}{\gamma_0} \quad \{\text{`` 1-}vs.\text{-2''}\}, \tag{20}$$

$$\frac{1}{\gamma_0}\mathrm{LR}_1 = \frac{\gamma_1}{\gamma_0} \quad \{\text{`` 1-}vs.\text{-3''}\}, \tag{21}$$

$$\frac{1}{\gamma_0}\mathrm{LR}_2 = \frac{\gamma_2}{\gamma_0} \quad \{\text{`` 2-}vs.\text{-3''}\}, \tag{22}$$

where $\gamma_0 \equiv \gamma_1 + \gamma_2 + 4$ (to impose consistency with (7)). Note the similarity in form between these equations and (8)–(10). If we require γ_1 and γ_2 to be positive, the correspondence is exact, and this special case of (8)–(10) is illustrated in Fig. 3. (In fact, the intersection of the ideal observer decision boundary lines can lie in any quadrant. However, given a set of decision boundary lines with slopes as depicted in Fig. 2, the occurrence of the intersection point in any quadrant other than the first would result in an ideal observer operating point for which no observations were assigned to class π_3. This "degenerate" case will not be considered here.) As an aside, it is of some interest to note that if $\gamma_1 = \gamma_2 = 1$, the decision boundary line equations reduce to $\mathrm{LR}_1 = \mathrm{LR}_2$, yielding $p(\vec{x}|\pi_1) = p(\vec{x}|\pi_2)$; $\mathrm{LR}_1 = 1$, yielding $p(\vec{x}|\pi_1) = p(\vec{x}|\pi_3)$; and $\mathrm{LR}_2 = 1$, yielding $p(\vec{x}|\pi_2) = p(\vec{x}|\pi_3)$. That is, the decision boundary lines correspond, in the observational data space, to the loci of intersection of the observational data probability density functions. (This is illustrated in Figs. 2B and 2C of Scurfield (1998).)

A second correspondence between Scurfield's decision rule and the ideal observer decision rule can be obtained by taking $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$, with γ_1 and γ_2 now unrestricted. Substituting this definition in

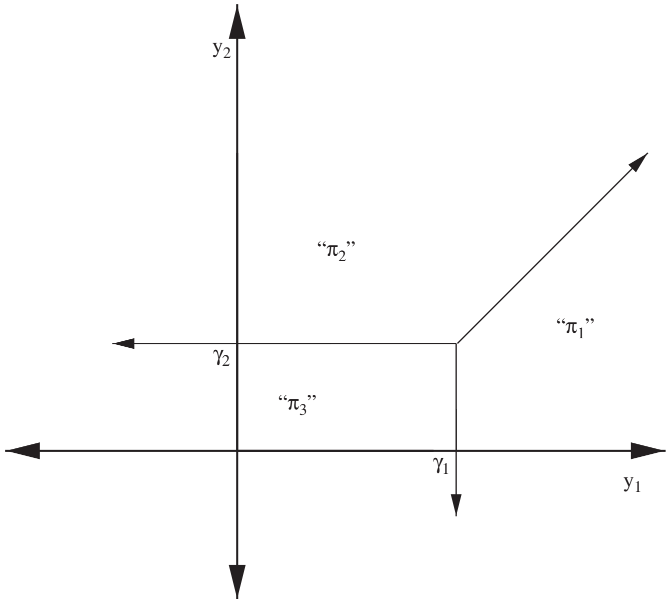


Fig. 2. Decision rule investigated by Scurfield, for the decision parameters γ_1 and γ_2.

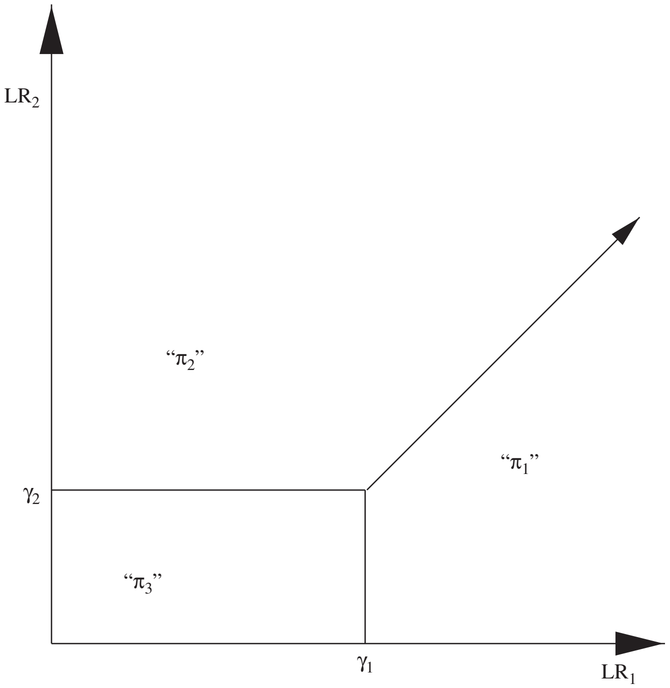


Fig. 3. A special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{212} = \gamma_{131} = \gamma_{232} = 1/(\gamma_1 + \gamma_2 + 4)$, $\gamma_{313} = \gamma_1/(\gamma_1 + \gamma_2 + 4)$, and $\gamma_{323} = \gamma_2/(\gamma_1 + \gamma_2 + 4)$. The parameters γ_1 and γ_2 are positive but otherwise arbitrary; this decision rule is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \mathrm{LR}_1(\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv \mathrm{LR}_2(\vec{\mathbf{x}})$.

(17)–(19), we obtain

$$\log(\mathrm{LR}_1) - \log(\mathrm{LR}_2) = \gamma_1 - \gamma_2 \quad \{\text{“ 1-}\mathit{vs.}\text{-2” }\}\ , \tag{23}$$

$$\log(\mathrm{LR}_1) = \gamma_1 \quad \{\text{“ 1-}\mathit{vs.}\text{-3” }\}\ , \tag{24}$$

$$\log(\mathrm{LR}_2) = \gamma_2 \quad \{\text{“ 2-}\mathit{vs.}\text{-3” }\}\ . \tag{25}$$

Taking exponentials on each side of these equations then gives

$$\frac{\mathrm{LR}_1}{\mathrm{LR}_2} = e^{\gamma_1-\gamma_2} \quad \{\text{“ 1-}\mathit{vs.}\text{-2” }\}\ , \tag{26}$$

$$\mathrm{LR}_1 = e^{\gamma_1} \quad \{\text{“ 1-}\mathit{vs.}\text{-3” }\}\ , \tag{27}$$

$$\mathrm{LR}_2 = e^{\gamma_2} \quad \{\text{“ 2-}\mathit{vs.}\text{-3” }\}\ , \tag{28}$$

we can then rearrange terms and divide the equations by a constant factor to obtain

$$\frac{e^{-\gamma_1}}{\gamma_0}\,\mathrm{LR}_1 - \frac{e^{-\gamma_2}}{\gamma_0}\,\mathrm{LR}_2 = 0 \quad \{\text{“ 1-}\mathit{vs.}\text{-2” }\}\ , \tag{29}$$

$$\frac{e^{-\gamma_1}}{\gamma_0}\,\mathrm{LR}_1 = \frac{1}{\gamma_0} \quad \{\text{“ 1-}\mathit{vs.}\text{-3” }\}\ , \tag{30}$$

$$\frac{e^{-\gamma_2}}{\gamma_0}\,\mathrm{LR}_2 = \frac{1}{\gamma_0} \quad \{\text{“ 2-}\mathit{vs.}\text{-3” }\}\ , \tag{31}$$

where $\gamma_0 \equiv 2(e^{-\gamma_1} + e^{-\gamma_2} + 1)$. By inspection, this is again a special case of (8)–(10), which is illustrated in Fig. 4. (This special case is currently the subject of independent analysis by He, Metz, Tsui, Links, & Frey, 2006.) As an aside, we note that if $\gamma_1 = \gamma_2 = 0$, the resulting decision boundary lines again correspond, in the observational data space, to the loci of intersection of the observational data probability density functions, as was pointed out in the text following (20)–(22).

Finally, if we take $\mathbf{y}_1 \equiv P(\pi_1|\vec{\mathbf{x}})$ and $\mathbf{y}_2 \equiv P(\pi_2|\vec{\mathbf{x}})$, and require $0<\gamma_1<1$ and $0<\gamma_2<1$, we obtain

$$P(\pi_1|\vec{x}) - P(\pi_2|\vec{x}) = \gamma_1 - \gamma_2 \quad \{\text{“ 1-}\mathit{vs.}\text{-2” }\}\ , \tag{32}$$

$$P(\pi_1|\vec{x}) = \gamma_1 \quad \{\text{“ 1-}\mathit{vs.}\text{-3” }\}\ , \tag{33}$$

$$P(\pi_2|\vec{x}) = \gamma_2 \quad \{\text{“ 2-}\mathit{vs.}\text{-3” }\}\ , \tag{34}$$

as illustrated in Fig. 5.

Note that (3) can be written as

$$\mathrm{LR}_i = \frac{P(\pi_i|\vec{x})p(\vec{x})/P(\pi_i)}{p(\vec{x}|\pi_3)} \qquad \{i : 1 \leqslant i \leqslant 2\},$$

$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i P(\pi_i)}{p(\vec{x})/p(\vec{x}|\pi_3)},$$

$$P(\pi_i|\vec{x}) = \frac{\mathrm{LR}_i[P(\pi_i)/P(\pi_3)]}{1 + \mathrm{LR}_1[P(\pi_1)/P(\pi_3)] + \mathrm{LR}_2[P(\pi_2)/P(\pi_3)]}. \tag{35}$$

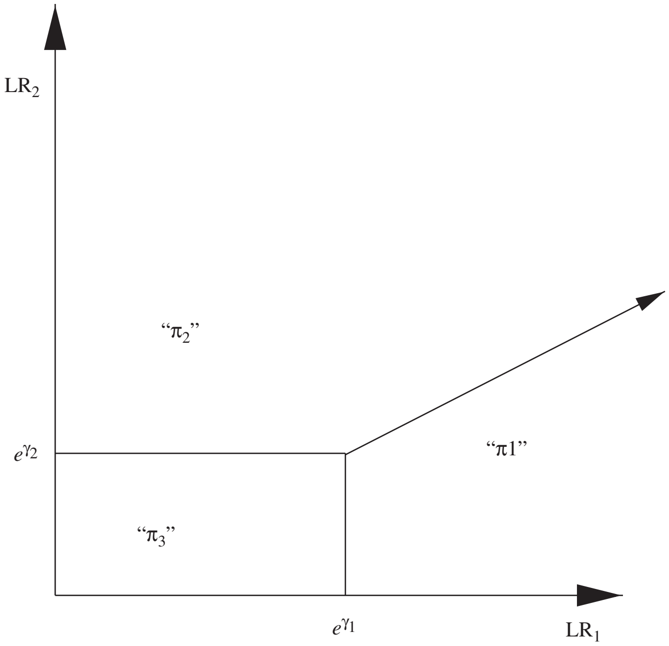


Fig. 4. A special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{131} = e^{-\gamma_1}/\gamma_0$, $\gamma_{212} = \gamma_{232} = e^{-\gamma_1}/\gamma_0$, $\gamma_{313} = \gamma_{323} = 1/\gamma_0$, and $\gamma_0 \equiv 2(e^{-\gamma_1} + e^{-\gamma_2} + 1)$. The parameters γ_1 and γ_2 are arbitrary; this decision rule is a special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$.

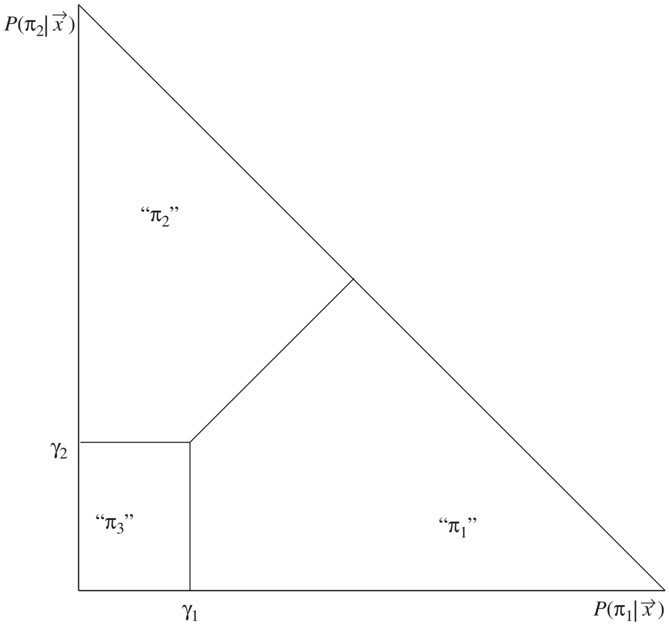


Fig. 5. A special case of the Scurfield decision rule with $\mathbf{y}_1 \equiv P(\pi_1|\vec{x})$ and $\mathbf{y}_2 \equiv P(\pi_2|\vec{x})$.

This allows us to rewrite (32)–(34) as

$$\frac{1-(\gamma_1-\gamma_2)}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\,\mathrm{LR}_1 - \frac{1+(\gamma_1-\gamma_2)}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_1-\gamma_2}{\gamma_0}, \tag{36}$$

$$\frac{1-\gamma_1}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 \quad -\frac{\gamma_1}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_1}{\gamma_0}, \tag{37}$$

$$-\frac{\gamma_2}{\gamma_0}\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 + \frac{1-\gamma_2}{\gamma_0}\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \frac{\gamma_2}{\gamma_0}, \tag{38}$$

respectively, where $\gamma_0 \equiv (2-2\gamma_1+\gamma_2)P(\pi_1)/P(\pi_3)+(2+\gamma_1-2\gamma_2)P(\pi_2)/P(\pi_3)+\gamma_1+\gamma_2$. This is again a special case of (8)–(10), as the quantities $1-(\gamma_1-\gamma_2)$, $1+(\gamma_1-\gamma_2)$, $1-\gamma_1$, and $1-\gamma_2$ are all positive given $0<\gamma_1<1$ and $0<\gamma_2<1$.

Scurfield (1998) points out that the observer which maximizes P_C, the "percent correct" or probability of a correct response, is a special case of the ideal observer (i.e., a single operating point achievable by the ideal observer for the given task). This observer follows the Scurfield decision rule model with $\mathbf{y}_1 \equiv \log(\mathrm{LR}_1(\vec{\mathbf{x}}))$ and $\mathbf{y}_2 \equiv \log(\mathrm{LR}_2(\vec{\mathbf{x}}))$, and decision parameters given by $e^{\gamma_1} = P(\pi_3)/P(\pi_1)$ and $e^{\gamma_2} = P(\pi_3)/P(\pi_2)$. It is interesting to note that the Scurfield decision rule model can in fact be used to describe ideal observer performance for an even wider class of operating points, as shown in this section.

To evaluate the performance of an observer using the decision rule in (17)–(19), Scurfield plots a set of six surfaces in three-dimensional ROC spaces, giving $P(\mathbf{d}=\pi_2|\mathbf{t}=\alpha(\pi_2))$ as a function of $P(\mathbf{d}=\pi_1|\mathbf{t}=\alpha(\pi_1))$ and $P(\mathbf{d}=\pi_3|\mathbf{t}=\alpha(\pi_3))$. Here α is one of the six possible permutations of three symbols. Scurfield gives a probabilistic interpretation for this evaluation methodology: the volume under each surface is the probability of a particular outcome in a three-alternative forced choice experiment, and thus the six volumes must sum to one. This constraint means that at most five of the surfaces are independent. However, given the number of conditional probabilities $P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j)$ involved, one can show that only four such surfaces are required to completely specify the tradeoffs among the observer's conditional classification probabilities. Without loss of generality, we consider plotting each of $P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_1)$, $P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_3)$, $P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_1)$, and $P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_2)$ as functions of $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_2)$ and $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_3)$. (As with Scurfield's plots, these are well defined because Scurfield's decision rule has two degrees of freedom, namely the parameters γ_1 and γ_2).

Now consider one of Scurfield's plots, for example that which gives $P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_2)$ as a function of $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_1)$ and $P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_3)$. Because these are conditional probabilities, we have

$$P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_1) = 1 - P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_1) - P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_1), \tag{39}$$

$$P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_2) = 1 - P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_2) - P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_2), \tag{40}$$

$$P(\mathbf{d}=\pi_3|\mathbf{t}=\pi_3) = 1 - P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_3) - P(\mathbf{d}=\pi_2|\mathbf{t}=\pi_3). \tag{41}$$

Each of the conditional probabilities on the right-hand side of these equations can be written as functions of $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_2)$ and $P(\mathbf{d}=\pi_1|\mathbf{t}=\pi_3)$ in our formulation; thus, the surface given in this plot is determined parametrically by the set of four surfaces we have given. Similar remarks hold for the other five surfaces used by Scurfield. In general, for an N-class classification task using a Scurfield-type decision rule with $N-1$ degrees of freedom (the generalization to N classes of (17)–(19)), one can show that a set of $(N-1)^2$ hypersurfaces with $N-1$ degrees of freedom in N-dimensional ROC spaces is necessary to fully characterize the observer's performance, although the interpretation of those hypersurfaces is not necessarily as straightforward or elegant as that provided for the $N!-1$ hypersurfaces used by Scurfield.

4. The Chan decision rule

Chan et al. are investigating three-class classifiers for computer-aided diagnosis (Chan et al., 2003). Their work is motivated by reasoning similar in principle to that which we independently arrived at when we began to consider this problem. In particular, they consider a clinical situation in which observations must be classified as malignant, benign, or normal. The goal of their work is not just the psychophysical measurement of the performance of an existing (e.g., human) observer, but the optimization of the performance of a system (containing components with parameters subject to experimental control, e.g. an artificial neural network) to aid a radiologist or clinician. Thus they are free, at least in theory, to start explicitly from an ideal observer model in constructing their decision rule.

In order to reduce the complexity of the ideal observer decision rule to manageable proportions, Chan et al. impose restrictions on the utilities used by their observer. In their formulation, the class we are labeling π_1 is the benign class; π_2, the normal class; and the malignant class is π_3. They further assume that the possible values of any utility $U_{i|j}$ are restricted to the interval $[0,1]$. They then set $U_{1|1}=U_{2|2}=U_{3|3}=1$ (i.e., correctly identifying any case has maximal utility). Furthermore, they require $U_{2|1}=U_{1|2}=1$ and $U_{1|3}=U_{2|3}=0$ (i.e., misidentifying a benign case as normal, or vice versa, has no significant cost reducing the utility of such a decision from the maximum, but misclassifying an actually malignant case as benign or normal has the minimum possible utility). Finally, $U_{3|1}$ and $U_{3|2}$ are assumed to have arbitrary values on the open interval $(0,1)$ (i.e., misclassifying an actually non-malignant case as malignant will have some cost reducing the utility of such a decision from the maximum, but such a misclassification is in some sense "better" than missing an actual malignancy). It is important to note that these assumptions are arguably relevant to a reasonable model of a clinical situation, and are thus of interest beyond their superficial advantage in reducing the degrees of freedom involved in the observer's decision rule. We will, however, only consider the latter issue in the remainder of this section.

Substituting the values of the utilities given above into (4), we obtain decision boundary lines of the form

$$0\,\mathrm{LR}_1 + 0\,\mathrm{LR}_2 = 0 \quad \{\text{“ 1-}vs.\text{-2” }\}, \tag{42}$$

$$\frac{(1-U_{3|1})P(\pi_1)}{\gamma_0}\,\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\pi_2)}{\gamma_0}\,\mathrm{LR}_2 = \frac{P(\pi_3)}{\gamma_0} \quad \{\text{“ 1-}vs.\text{-3” }\}, \tag{43}$$

$$\frac{(1-U_{3|1})P(\pi_1)}{\gamma_0}\mathrm{LR}_1 + \frac{(1-U_{3|2})P(\pi_2)}{\gamma_0}\mathrm{LR}_2 = \frac{P(\pi_3)}{\gamma_0} \quad \{\text{“ 2-}vs.\text{-3” }\}, \tag{44}$$

where $\gamma_0 \equiv 1 + P(\pi_3) - U_{3|1}P(\pi_1) - U_{3|2}P(\pi_2)$. Note that, as Chan et al. point out, the “1-*vs.*-2” line is in fact undefined for this choice of utilities, while the “1-*vs.*-3” and “2-*vs.*-3” lines are identical. This is a general consequence of (8)–(10); if any two of these equations yield identical lines, the third line must be undefined. (Note that, strictly speaking, the utility structure employed by Chan et al. is excluded from our formulation by the requirement stated in (6). However, this issue—i.e., whether the ideal observer's performance should be considered to include such limiting cases—is largely a definitional, rather than a fundamental, issue, because (6) could just as readily have been formulated as a non-negativity constraint, rather than a strict inequality as we have chosen).

The decision rule considered by Chan et al. is illustrated in Fig. 6. It can be argued that, in a sense, the output of this classifier belongs to only two classes, malignant and non-malignant; in particular, because (42) is undefined, this observer will never unequivocally decide $\mathbf{d} = \pi_1$ (benign) or π_2 (normal). In fact, if $U_{3|1} = U_{3|2}$, the observer's performance is identical with that of a two-class ideal observer which distinguishes between the malignant and non-malignant (benign plus normal) classes. However, in the more general case in which $U_{3|1} \neq U_{3|2}$, the observer considered by Chan et al. is able to achieve ROC operating points not accessible by the two-class ideal observer. (That is, the three-class ideal observer can achieve points below the two-class ideal observer's ROC curve in a two-class ROC space, or, equivalently, points off the curve representing the two-class ideal observer's performance plotted in a three-class ROC space.) Intuitively, their observer makes decisions based on the three distribution functions of the observational data, even though the observer's output consists of only two possible responses.

Chan et al. evaluate the performance of their observer by plotting $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3)$ as a function of $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_1)$ and $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_2)$. Note that this single two-dimensional surface is sufficient to completely characterize the tradeoffs among the conditional classification probabilities of their observer. This is because, as just stated, the observer's output consists of only two possible responses, and thus we have only six classification probabilities $P(\mathbf{d} = \pi_i|\mathbf{t} = \pi_j)$ rather than the nine expected in a three-class classification task. These six conditional probabilities are still constrained by three equations, however:

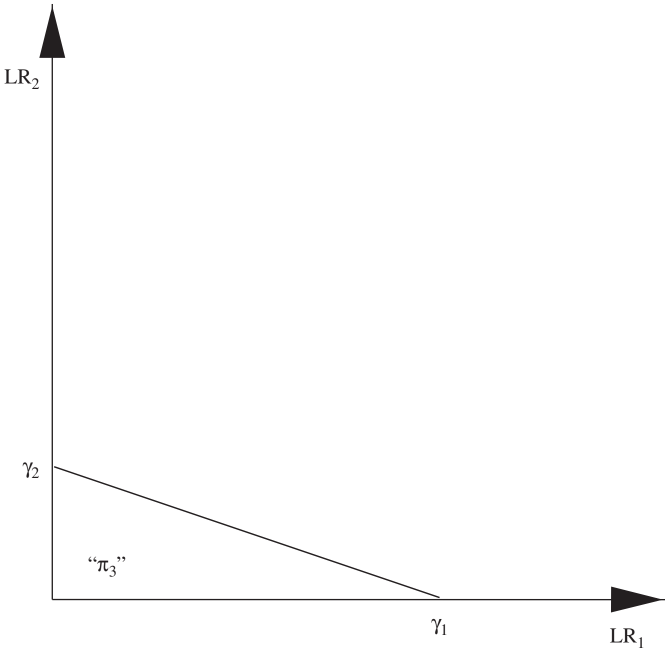


Fig. 6. The decision rule investigated by Chan et al., which is a special case of the ideal observer decision rule with $\gamma_{121} = \gamma_{212} = 0$, $\gamma_{131} = (1-U_{3|1})P(\pi_1)/\gamma_0$, $\gamma_{232} = (1-U_{3|2})P(\pi_2)/\gamma_0$, and $\gamma_{313} = \gamma_{323} = P(\pi_3)/\gamma_0$; here $\gamma_0 \equiv 1 + P(\pi_3) - U_{3|1}P(\pi_1) - U_{3|2}P(\pi_2)$. Observations in the unlabeled region are decided “not π_3”, i.e., either “π_1” or “π_2”. The intercepts γ_1 and γ_2 are $P(\pi_3)/[(1-U_{3|1})P(\pi_1)]$ and $P(\pi_3)/[(1-U_{3|2})P(\pi_2)]$, respectively.

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_1) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_1) = 1, \tag{45}$$

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_2) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_2) = 1, \tag{46}$$

$$P(\mathbf{d} = \tilde{\pi}_3|\mathbf{t} = \pi_3) + P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3) = 1, \tag{47}$$

where the expression $\mathbf{d} = \tilde{\pi}_3$ indicates that the observer decides that the observation does not belong to class π_3. These constraint equations allow us to eliminate three of the six conditional probabilities, leaving a single ROC surface with two degrees of freedom in a three-dimensional ROC space.

5. The Mossman decision rule

Mossman (1999) investigates a decision rule applied to a set of three decision variables $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$, subject to the constraint

$$\mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 = 1, \tag{48}$$

as well as $0 \leqslant \mathbf{y}_i \leqslant 1$ $\{1 \leqslant i \leqslant 3\}$. This is consistent with the constraint on the a posteriori class probabilities, $P(\pi_1|\vec{\mathbf{x}}) + P(\pi_2|\vec{\mathbf{x}}) + P(\pi_3|\vec{\mathbf{x}}) = 1$; these quantities are known to be directly related to the likelihood ratio ideal observer decision variables (Edwards, Lan, Metz, Giger, &

Nishikawa, 2004a; Kupinski, Edwards, Giger, & Metz, 2001). Mossman does not explicitly require, however, that the decision variables in (48) be the a posteriori class probabilities (e.g., they may be noisy estimates of these quantities).

The decision rule considered by Mossman, which depends on two decision parameters γ_1 and γ_2, is

$$\text{decide} \quad d = \pi_1 \quad \text{iff } y_2 - y_1 \leqslant \gamma_2 \quad \text{and} \quad y_3 \leqslant \gamma_1, \tag{49}$$

$$\text{decide} \quad d = \pi_2 \quad \text{iff } y_2 - y_1 > \gamma_2 \quad \text{and} \quad y_3 \leqslant \gamma_1, \tag{50}$$

$$\text{decide} \quad d = \pi_3 \quad \text{iff } y_3 > \gamma_1. \tag{51}$$

where $0 \leqslant \gamma_1 \leqslant 1$ and $-1 \leqslant \gamma_2 \leqslant 1$. From these relations, and given the relation $y_3 = 1 - y_1 - y_2$ from (48), one can define the decision boundary lines

$$y_1 - y_2 = -\gamma_2 \quad \{\text{“ 1-}vs.\text{-2” }\}, \tag{52}$$

$$y_1 + y_2 = 1 - \gamma_1 \quad \{\text{“ 1-}vs.\text{-3” }\}, \tag{53}$$

$$y_1 + y_2 = 1 - \gamma_1 \quad \{\text{“ 2-}vs.\text{-3” }\}. \tag{54}$$

This decision rule is illustrated in Fig. 7. Note that, similar to the Chan et al. decision rule, the “1-*vs.*-3” and “2-*vs.*-3” decision boundary lines are identical.

We now consider a special case of the Mossman decision rule in which $\mathbf{y}_1 = P(\pi_1|\vec{\mathbf{x}})$, $\mathbf{y}_2 = P(\pi_2|\vec{\mathbf{x}})$, and $\mathbf{y}_3 = P(\pi_3|\vec{\mathbf{x}})$ for some observational data vector $\vec{\mathbf{x}}$. As in Section 3, we make the substitution in (35); this allows us to rewrite (52)–(54) as

$$(1+\gamma_2)\frac{P(\pi_1)}{P(\pi_3)}\,\mathrm{LR}_1 - (1-\gamma_2)\frac{P(\pi_2)}{P(\pi_3)}\,\mathrm{LR}_2 = -\gamma_2 \quad \{\text{“ 1-}vs.\text{-2” }\}, \tag{55}$$

$$\gamma_1\frac{P(\pi_1)}{P(\pi_3)}\,\mathrm{LR}_1 + \gamma_1\frac{P(\pi_2)}{P(\pi_3)}\,\mathrm{LR}_2 = 1 - \gamma_1 \quad \{\text{“ 1-}vs.\text{-3” }\}, \tag{56}$$

$$\gamma_1\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 + \gamma_1\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = 1 - \gamma_1 \quad \{\text{“ 2-}vs.\text{-3” }\}, \tag{57}$$

This version of the decision rule is illustrated in Fig. 8.

Although the Mossman decision rule for this choice of decision variables appears similar in form to the ideal observer decision rule, recall from Section 4 that if two of the decision boundary line equations are identical, the third must yield a line identical to the first two or be undefined. Another way to see this is to note that the coefficients of (10) are differences of the corresponding coefficients of (8) and (9). If the coefficients of (9) and (10) are identical, it must be the case that the coefficients of (8) are all zero. For the Mossman decision rule, this would require $1 + \gamma_2 = 0$, $1 - \gamma_2 = 0$, and $\gamma_2 = 0$ simultaneously, which is clearly impossible.

It follows that, for this particular choice of decision variables (related in a straightforward way to the ideal

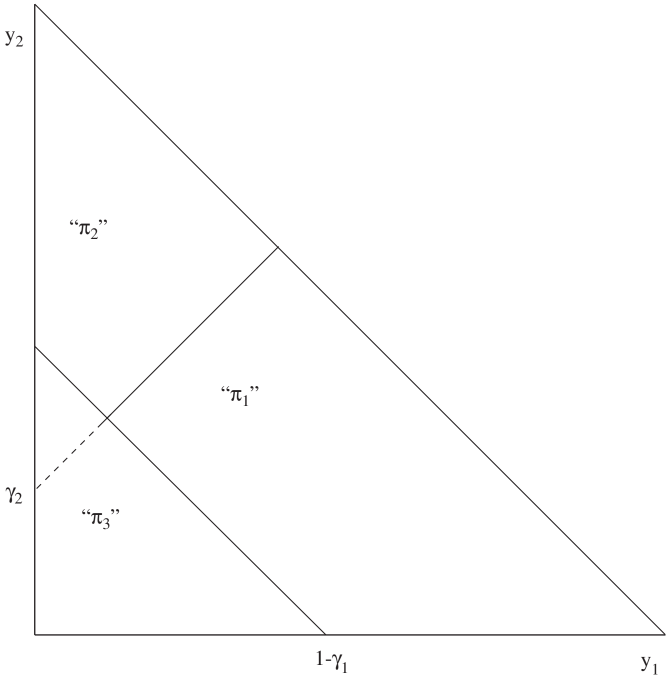


Fig. 7. Decision rule investigated by Mossman, for the decision parameters γ_1 and γ_2, shown in the a posteriori class probability space.

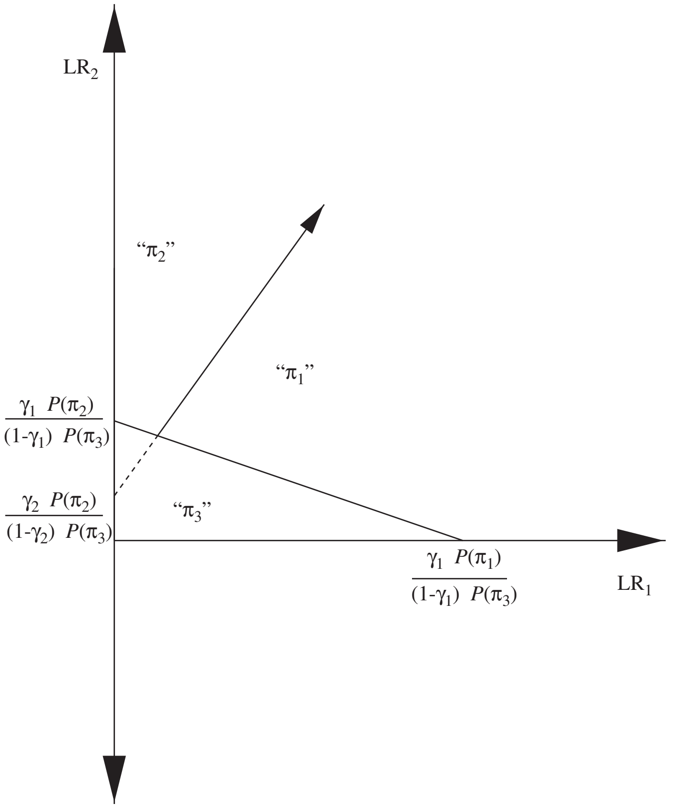

LR_2
“π_2”
“π_1”
$\frac{\gamma_1 \; P(\pi_2)}{(1-\gamma_1) \; P(\pi_3)}$
$\frac{\gamma_2 \; P(\pi_2)}{(1-\gamma_2) \; P(\pi_3)}$
“π_3”
$\frac{\gamma_1 \; P(\pi_1)}{(1-\gamma_1) \; P(\pi_3)}$
LR_1

Fig. 8. Decision rule investigated by Mossman, for the decision parameters γ_1 and γ_2, shown in likelihood ratio space.

observer's decision variables), the decision rule considered by Mossman cannot represent possible ideal observer performance for any choice of the utilities $U_{i|j}$ in (1) and (2). (One can construct probability density functions such that the Mossman observer's behavior for a particular choice of decision criteria (γ_1 and γ_2 in (49)–(51)) corresponds to ideal observer behavior at a particular operating point. However, we do not at present have any reason to believe that this result can be generalized to arbitrary probability density functions or to arbitrary choices of decision criteria for a given choice of probability density functions).

Mossman proposed that the ROC surface obtained by plotting $P(\mathbf{d} = \pi_3|\mathbf{t} = \pi_3)$ as a function of $P(\mathbf{d} = \pi_1|\mathbf{t} = \pi_1)$ and $P(\mathbf{d} = \pi_2|\mathbf{t} = \pi_2)$ be used to evaluate the performance of the observer. Although this surface is clearly well-defined (the Mossman decision rule has two degrees of freedom, namely the parameters γ_1 and γ_2), it follows from the discussion at the end of Section 3 that four such surfaces in three-dimensional ROC spaces are needed to completely characterize the tradeoffs among the observer's conditional classification probabilities.

6. Discussion and conclusions

We examined three decision rules proposed recently for three-class classification tasks by different researchers. The basis for our evaluation was ideal observer decision theory, primarily because our own interest in the three-class classification task is its possible application to CAD. A major goal in the development of a computerized scheme for CAD is the optimization of the performance of that scheme, in order to provide the maximum benefit to clinicians and thus to their patients. It should thus be kept clearly in mind that the ideal observer framework may not be as relevant, for example, to work which is motivated by purely psychophysical considerations (Mossman, 1999; Scurfield, 1996, 1998)—i.e., where the goal is to estimate of the properties of an existing observer.

That being said, the three-class classification task is difficult enough that it is perhaps worth making any attempt to analyze, from a single point of view, the work of the relatively few researchers investigating this problem, even in cases where that point of view is not necessarily relevant to the underlying motivations for that work. We feel the insights we have gained from the analysis of various decision rules presented here should provide at least some justification for that claim.

In particular, Scurfield points out (Scurfield, 1998) that his proposed decision rule is in fact an ideal observer decision rule for a single ideal observer operating point, namely the observer which maximizes the probability of any correct response (or "percent correct" or P_C). We were able to show that, under various assumptions, a larger set of such correspondences between the Scurfield observer and the ideal observer exists.

Chan et al. (2003) are working on the application of three-class classification to CAD, and thus explicitly take the ideal observer as the starting point in the development of their decision rule. Although this rendered our analysis of that decision rule in terms of ideal observer decision theory largely trivial, their decision rule merits attention as an example of a situation in which the ideal observer is indeed making use of information from the three classes of observations (i.e., its behavior is demonstrably different from that of a two-class ideal observer), while only producing two different responses for those observations. In two-class classification, the only corresponding examples are trivial: either the observer always calls observations positive (achieving an operating point of $(\mathrm{FPF} = 1, \mathrm{TPF} = 1)$, where FPF is the false-positive fraction and TPF the true-positive fraction) or always calls them negative $(\mathrm{FPF} = 0, \mathrm{TPF} = 0)$.

Finally, we showed that, given a particular and obvious choice of ideal-observer-related decision variables, the decision rule proposed by Mossman (1999) does not correspond to ideal observer behavior for any possible values of the observer's utilities. However, we note that the structure of the Mossman decision rule—a simple sequence of thresholds on single decision variables—may indeed serve as a reasonable model for human observer performance in certain situations, e.g., differential diagnosis. That such a decision rule fails to be an ideal observer decision rule may be considered surprising, given the properties the Mossman decision rule shares with that of Chan et al.—in particular, the identity of two out of the three decision boundary lines. The reasons why one decision rule can be said to correspond to ideal observer behavior, while a rule similar in structure does not when used with a particular and obvious choice of decision variables, are connected to fundamental constraints on the ideal observer's behavior; given the inherent complexities of the three-class classification task, it is easy for such subtleties to be overwhelmed by other details. A close comparison of two possible three-class classification decision rules can thus provide an immediate and intuitive understanding of such properties, even though a complete and fully general solution to the three-class classification problem remains elusive.

Acknowledgments

The authors thank Vit Drga for bringing Brian Scurfield's work to their attention, and thank both Vit Drga and Brian Scurfield for helpful conversations concerning that work. The authors thank Heang-Ping Chan and Berkman Sahiner for helpful conversations concerning their work.

References

Bick, U., Giger, M. L., Schmidt, R. A., Nishikawa, R. M., Wolverton, D. E., & Doi, K. (1995). Automated segmentation of digitized mammograms. *Academic Radiology*, *2*, 1–9.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., & Simmons, A. H. (1978). A free response approach to the measurement and characterization of radiographic-observer performance. *Journal of Applied Photographic Engineering*, *4*, 166–172.

Chakraborty, D. P. (1989). Maximum likelihood analysis of free-response operating characteristic (FROC) data. *Medical Physics*, *16*, 561–568.

Chakraborty, D. P. (2002). Statistical power in observer-performance studies: Comparison of the receiver operating characteristic and free-response methods in tasks involving localization. *Academic Radiology*, *9*, 147–156.

Chan, H.-P., Sahiner, B., Hadjiiski, L. M., Petrick, N., & Zhou, C. (2003). Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study. In M. Sonka, J.M. Fitzpatrick (Eds.), *Proceedings of the SPIE*, *medical imaging 2003*: *Image processing* (Vol. 5032, pp. 567–578). Bellingham, WA: SPIE.

Dreiseitl, S., Ohno-Machado, L., & Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, *20*, 323–331.

Edwards, D. C., Lan, L., Metz, C. E., Giger, M. L., & Nishikawa, R. M. (2004a). Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions. *Medical Physics*, *31*, 81–90.

Edwards, D. C., Metz, C. E., & Kupinski, M. A. (2004b). Ideal observers and optimal ROC hypersurfaces in *N*-class classification. *IEEE Transactions on Medical Imaging*, *23*, 891–895.

Edwards, D. C., Metz, C. E., & Nishikawa, R. M. (2005). The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in *N*-class classification tasks. *IEEE Transactions on Medical Imaging*, *24*, 293–299.

Ferri, C., Hernández-Orallo, J., & Salido, M. A. (2003). *Volume under the ROC surface for multi-class problems: Exact computation and evaluation of approximations*. Technical Report, Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València (Spain).

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*, 171–186.

He, X., Metz, C. E., Tsui, B. M. W., Links, J. M., & Frey, E. C. (2006). Three-class ROC analysis—A decision theoretic approach under the ideal observer framework. *IEEE Transactions on Medical Imaging*, *25*, 571–581.

Huo, Z., Giger, M. L., & Metz, C. E. (1999). Effect of dominant features on neural network performance in the classification of mammographic lesions. *Physics in Medicine and Biology*, *44*, 2579–2595.

Huo, Z., Giger, M. L., & Vyborny, C. J. (2001). Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, *20*, 1285–1292.

Huo, Z., Giger, M. L., Vyborny, C. J., & Metz, C. E. (2002). Breast cancer: Effectiveness of computer-aided diagnosis—Observer study with independent database of mammograms. *Radiology*, *224*, 560–568.

Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., & Metz, C. E. (2000). Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness. *Academic Radiology*, *7*, 1077–1084.

Huo, Z., Giger, M. L., Vyborny, C. J., Wolverton, D. E., Schmidt, R. A., & Doi, K. (1998). Automated computerized classification of malignant and benign masses on digitized mammograms. *Academic Radiology*, *5*, 155–168.

Kupinski, M. A. (2000). *Computerized pattern classification in medical imaging*. Ph.D. Thesis, The University of Chicago, Chicago, IL.

Kupinski, M. A., Edwards, D. C., Giger, M. L., & Metz, C. E. (2001). Ideal observer approximation using Bayesian classification neural networks. *IEEE Transactions on Medical Imaging*, *20*, 886–899.

Lachiche, N., & Flach, P. (2003). Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In *Proceedings of the twentieth international conference on machine learning (ICML-2003)* (pp. 416–423). Washington, DC: AAAI Press.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, *19*, 78–89.

Nakas, C. T., & Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, *23*, 3437–3449.

Obuchowski, N. A., Applegate, K. E., Goske, M. J., Arheart, K. L., Myers, M. T., & Morrison, S. (2001). The 'differential diagnosis' for multiple diseases: Comparison with the binary-truth state experiment in two empirical studies. *Academic Radiology*, *8*, 947–954.

Patel, A. C., Markey, M. K. (2005). Comparison of three-class classification performance metrics: A case study in breast cancer CAD. In M.P. Eckstein, Y. Jiang (Eds.), *Proceedings of the SPIE*, *medical imaging 2005*: *Image perception*, *observer performance*, *and technology assessment* (Vol. 5749, pp. 581–589). Bellingham, WA, SPIE.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, *40*, 253–269.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability to *n*-event *m*-dimensional forced-choice tasks. *Journal of Mathematical Psychology*, *42*, 5–31.

Srinivasan, A. (1999). *Note on the location of optimal classifiers in n-dimensional ROC space*. Technical Report PRG-TR-2-99, Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford.

Van Trees, H. L. (1968). *Detection, estimation and modulation theory: Part I*. New York: Wiley.

Yin, F.-F., Giger, M. L., Doi, K., Metz, C. E., Vyborny, C. J., & Schmidt, R. A. (1991). Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images. *Medical Physics*, *18*, 955–963.

Yin, F.-F., Giger, M. L., Doi, K., Vyborny, C. J., & Schmidt, R. A. (1994). Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique. *Medical Physics*, *21*, 445–452.

Yin, F.-F., Giger, M. L., Vyborny, C. J., Doi, K., & Schmidt, R. A. (1993). Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses. *Investigation Radiology*, *28*, 473–481.

G Restrictions on the Three-Class Ideal Observer's Decision Boundary Lines

Restrictions on the Three-Class Ideal Observer's Decision Boundary Lines

Darrin C. Edwards* and Charles E. Metz

***Abstract*—We are attempting to develop expressions for the coordinates of points on the three-class ideal observer's receiver operating characteristic (ROC) hypersurface as functions of the set of decision criteria used by the ideal observer. This is considerably more difficult than in the two-class classification task, because the conditional probabilities in question are not simply related to the cumulative distribution functions of the decision variables, and because the slopes and intercepts of the decision boundary lines are not independent; given the locations of two of the lines, the location of the third will be constrained depending on the other two. In this paper, we attempt to characterize those constraining relationships among the three-class ideal observer's decision boundary lines. As a result, we show that the relationship between the decision criteria and the misclassification probabilities is not one-to-one, as it is for the two-class ideal observer.**

***Index Terms*—Ideal observers, ROC analysis, three-class classification.**

I. Introduction

Receiver operating characteristic (ROC) analysis is the accepted methodology for analyzing the performance of a two-class classifier [1], in particular for medical decision-making tasks in which a patient is diagnosed as having or not having a particular condition based on features of a medical image [2]. In judging the performance of an observer measured via ROC analysis, the standard for comparison is the so-called ideal observer, that observer which outperforms any other possible observer given the statistical variability of the observational data being classified [1], [3]. Although the general form of the ideal observer in a classification task with three or more classes has been known for some time [3], the considerable complexities inherent to this model compared to the two-class classification task have hampered the development of extensions of ROC analysis which are both fully general and practically useful. (Several researchers have recently proposed restricted observer models or restricted evaluation methods [4]–[7].)

Despite these difficulties, research continues in this area because the advantages to be gained from a three-class classifier and appropriate evaluation methodology are considerable. In our own case, we seek to combine existing computer-aided diagnosis (CAD) schemes for detecting [8]–[12] mammographic mass lesions and classifying [13]–[17] them as malignant or benign. The combined scheme would serve as a fully automated classifier (the existing classifier requires initial manual identification of lesions by a radiologist), potentially allowing radiologists to reduce their false-positive biopsy rate without reducing their sensitivity for detection of malignancies. Simply concatenating the two types of scheme in a two-stage classifier would be inadequate, because the output of the detection scheme will necessarily include false-positive (FP) computer detections in addition to the malignant and benign lesions to be classified. These FP computer detections correspond to objects which were by design not included in the training sample of the classification scheme, because they are not members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme's output cannot be used unmodified as the input to the classification scheme.

Our initial efforts toward the goal of developing a true three-class classifier have been more theoretical than practical so far. We have shown that, just as the two-class ideal observer achieves the optimal two-class ROC curve for a given task, the N-class ideal observer achieves the optimal N-class ROC hypersurface [18]. (Note that the ideal observer is formally defined as that which minimizes the expected Bayes risk [3], and not in terms of classification performance, making this a nontrivial observation in both cases.) More soberingly, we found recently that an obvious generalization of the well-known performance metric, the area under the ROC curve (AUC), is not a useful performance metric in a classification task with three or more classes [19].

At present we are attempting to develop expressions for the coordinates of points on the three-class ideal observer's ROC hypersurface (the conditional probabilities for misclassifying observations [18], [20], [21]) as functions of the set of decision criteria used by the ideal observer. This is considerably more difficult than in the two-class classification task for two reasons. First, the conditional probabilities in question are not simply related to the cumulative distribution functions (cdfs) of the decision variables, but are integrals of those variables over domains determined by three decision boundary lines [3]. Second, the slopes and intercepts of the decision boundary lines are not independent; given the locations of two of the lines, we have found recently that the location of the third will be constrained depending on the other two.

In this paper, we attempt to characterize the constraining relationships just mentioned among the three-class ideal observer's

Manuscript received May 17, 2005; revised August 25, 2005. This work was supported by the U.S. Army Medical Research and Materiel Command under Grant W81XWH-04-1-0495 (D. C. Edwards, principal investigator). The Associate Editor responsible for coordinating the review of this paper and recommending its publication was E. Krupinski. *Asterisk indicates corresponding author.*

*D. C. Edwards is with the Department of Radiology, the University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637 USA (e-mail: d-edwards@uchicago.edu).

C. E. Metz is with the Department of Radiology, the University of Chicago, Chicago, IL 60637 USA.

Digital Object Identifier 10.1109/TMI.2005.859212

decision boundary lines. Although this paper is admittedly still removed from image analysis perse, we hope it may prove of interest to the CAD community and ultimately of relevance to a wide variety of medical image analysis tasks. In the next section we briefly review the structure of the three-class ideal observer and the notation we have been using to characterize it [18]. In Section III, we show that for a given location (slope and y-intercept) of the decision boundary line separating the first and third classes, the location of one of the remaining two lines is constrained in a particular way based on the location of the other.

These results are discussed in Section IV. Given the arbitrariness of the labels applied to the three classes (ie, which classes are considered first, second, or third), one would expect the selection of the fixed line in Section III to be similarly arbitrary, and indeed in Appendices A and B we show that corresponding and consistent results are obtained if one takes the location of the decision boundary line separating the second and third, or first and second, classes, respectively, to be given.

II. The Three-Class Ideal Observer

In [18], we showed that an N-class ideal observer makes decisions by partitioning a likelihood ratio decision variable space, where the boundaries of the partitions are given by hyperplanes

$$\text{decide}\quad d=\pi_i \text{ iff } \sum_{k=1}^{N-1}(U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k)\mathrm{LR}_k \geq (U_{j|N}-U_{i|N})P(\mathbf{t}=\pi_N) \quad \{j<i\} \tag{1}$$

$$\text{and } \sum_{k=1}^{N-1}(U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k)\mathrm{LR}_k > (U_{j|N}-U_{i|N})P(\mathbf{t}=\pi_N) \quad \{j>i\}. \tag{2}$$

Here, $U_{i|j}$ is the utility of deciding an observation is from class π_i given that it is actually from class π_j; $P(\mathbf{t}=\pi_k)$ is the apriori probability that an observation is drawn from class π_k; and LR_k is the kth likelihood ratio, defined by the ratio $p(\vec{x}|\pi_k)/p(\vec{x}|\pi_N)$ of the probability density functions of the observational data (We use boldface type to denote random variables). The partitioning is determined by the parameters

$$\gamma_{ijk}\equiv(U_{i|k}-U_{j|k})P(\mathbf{t}=\pi_k) \tag{3}$$

with i, j, and k varying from 1 to N, and $j\neq i$. Note that these parameters are not independent, however, because

$$\gamma_{ijk}=\gamma_{kjk}-\gamma_{kik}. \tag{4}$$

We can impose the reasonable condition that the utility for correctly classifying an observation from a given class should be greater than any utility for incorrectly classifying an observation from the same class, i.e., $U_{i|i}>U_{j|i}$ $\{i\neq j\}$. This gives, for $j\neq i$,

$$\gamma_{iji}>0 \tag{5}$$

leaving $N(N-1)$ positive parameters (the rest are derivable from (4)).

Finally, note that the hyperplanes represented by (1) and (2) are unchanged if we multiply all of these equations by a single scalar, such as $1/(\Sigma_{i\neq j}\gamma_{iji})$. This leaves us with N^2-N-1 degrees of freedom, as expected.

The behavior of a three-class ideal observer is completely determined by the three decision boundary lines

$$\gamma_{121}\mathrm{LR}_1-\gamma_{212}\mathrm{LR}_2=\gamma_{313}-\gamma_{323} \tag{6}$$

$$\gamma_{131}\mathrm{LR}_1+(\gamma_{232}-\gamma_{212})\mathrm{LR}_2=\gamma_{313} \tag{7}$$

$$(\gamma_{131}-\gamma_{121})\mathrm{LR}_1+\gamma_{232}\mathrm{LR}_2=\gamma_{323} \tag{8}$$

which we call, respectively, the "1-vs-2" line, the "1-vs-3" line, and the "2-vs-3" line. Note that if any two of these lines intersect, the third line must also share this intersection point. We also emphasize the simple interpretation, from (3), of each of the γ_{iji} parameters appearing in these decision boundary line equations as the difference in utilities between a "correct" and one particular "incorrect" decision (scaled by the apriori probability of the true class in question); and of each difference in the γ_{iji} parameters as a difference in utilities between two possible "incorrect" decisions [again scaled by the apriori probability of the true class in question; e.g., $\gamma_{313}-\gamma_{323}=(U_{2|3}-U_{1|3})P(\mathbf{t}=\pi_3)$].

From the conditions on the γ_{iji} parameters in (5), we can readily derive conditions on the decision boundaries themselves. If we denote the slope of the "i-vs-j" line by m_{ij}, its y-intercept by b_{ij}, and its x-intercept by χ_{ij}, we have

$$m_{12}=\frac{\gamma_{121}}{\gamma_{212}}>0 \tag{9}$$

$$\chi_{13}=\frac{\gamma_{313}}{\gamma_{131}}>0 \tag{10}$$

$$b_{23}=\frac{\gamma_{323}}{\gamma_{232}}>0. \tag{11}$$

These are the three conditions stated in [22].

III. Restrictions Determined by the Parameters of the "1-vs.-3" Line

Constraints on the decision boundaries, in addition to those given in (9)–(11), can be obtained by considering the two cases $\gamma_{232}-\gamma_{212}>0$ and $\gamma_{232}-\gamma_{212}<0$. In the first case (ie, $\gamma_{232}>\gamma_{212}$, or $U_{1|2}>U_{3|2}$), we have

$$m_{13}=\frac{-\gamma_{131}}{\gamma_{232}-\gamma_{212}}<0 \tag{12}$$

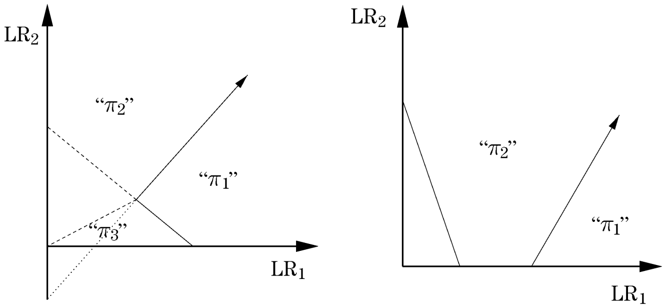
$$b_{13}=\frac{\gamma_{313}}{\gamma_{232}-\gamma_{212}}>0. \tag{13}$$

We also have

$$\begin{aligned} m_{23}&=\frac{-(\gamma_{131}-\gamma_{121})}{\gamma_{232}}\\ &=\frac{(\gamma_{232}-\gamma_{212})m_{13}+\gamma_{212}m_{12}}{\gamma_{232}}\\ &=\left(1-\frac{\gamma_{212}}{\gamma_{232}}\right)m_{13}+\frac{\gamma_{212}}{\gamma_{232}}m_{12}. \end{aligned} \tag{14}$$

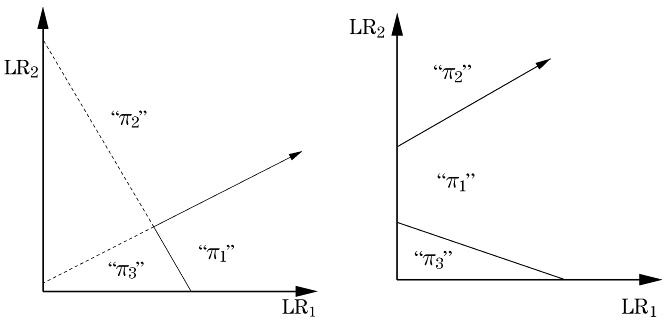
This is a weighted sum of the slopes m_{12} and m_{13}, where the weights are positive and sum to one. Since we must have $m_{13}<m_{12}$ from (9) and (12), it must therefore be the case that

$$m_{13}\leq m_{23}\leq m_{12}. \tag{15}$$



(a) (b)

Fig. 1. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} > 0$ (implying $m_{13} < 0$ and $b_{13} > 0$) and $b_{12} < 0$. In (a), $\chi_{12} < \chi_{13}$, and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $b_{23} > 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $\chi_{12} \geq \chi_{13}$ and the "2-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "1-vs-3" lines shown); observations above this line will be decided "π_2," and those below this line will be decided "π_3."



(a) (b)

Fig. 2. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} > 0$ (implying $m_{13} < 0$ and $b_{13} > 0$) and $b_{12} \geq 0$. In (a), $b_{12} < b_{13}$, and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $b_{12} \geq b_{13}$ and the "2-vs-3" line can lie anywhere between the "1-vs-2" and "1-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_1" regardless of the position of this line.

Furthermore

$$\begin{aligned} b_{23} &= \frac{\gamma_{323}}{\gamma_{232}} \\ &= \frac{\gamma_{313} - (\gamma_{313} - \gamma_{323})}{\gamma_{232}} \\ &= \frac{(\gamma_{232} - \gamma_{212})b_{13} + \gamma_{212}b_{12}}{\gamma_{232}} \\ &= \left(1 - \frac{\gamma_{212}}{\gamma_{232}}\right) b_{13} + \frac{\gamma_{212}}{\gamma_{232}} b_{12}. \end{aligned} \tag{16}$$

This is a weighted sum of the y-intercepts b_{12} and b_{13}, where the weights are positive and sum to one; thus, in addition to (15), we have the condition

$$\min(b_{12}, b_{13}) \leq b_{23} \leq \max(b_{12}, b_{13}). \tag{17}$$

If $b_{12} < 0$, then (17) immediately reduces to $b_{12} \leq b_{23} \leq b_{13}$ (by (13), we are considering a special case in which $b_{13} > 0$). This is illustrated in Fig. 1 for the slightly different situations $\chi_{12} < \chi_{13}$ and $\chi_{12} \geq \chi_{13}$. If, on the other hand, $b_{12} \geq 0$, then (15) and (17) together imply two possible situations, depending on whether $b_{12} < b_{13}$ or $b_{12} \geq b_{13}$. These possibilities are illustrated in Fig. 2.

We now consider the case $\gamma_{232} - \gamma_{212} < 0$ (ie, $\gamma_{232} < \gamma_{212}$, or $U_{1|2} < U_{3|2}$), which yields

$$m_{13} = \frac{-\gamma_{131}}{\gamma_{232} - \gamma_{212}} > 0 \tag{18}$$

$$b_{13} = \frac{\gamma_{313}}{\gamma_{232} - \gamma_{212}} < 0. \tag{19}$$

We now have

$$\begin{aligned} m_{12} &= \frac{\gamma_{121}}{\gamma_{212}} \\ &= \frac{\gamma_{131} - (\gamma_{131} - \gamma_{121})}{\gamma_{212}} \\ &= \frac{-(\gamma_{232} - \gamma_{212})m_{13} + \gamma_{232}m_{23}}{\gamma_{212}} \\ &= \left(1 - \frac{\gamma_{232}}{\gamma_{212}}\right) m_{13} + \frac{\gamma_{232}}{\gamma_{212}} m_{23}. \end{aligned} \tag{20}$$

This is again a weighted sum in which the weights are positive and sum to one, giving

$$\min(m_{13}, m_{23}) \leq m_{12} \leq \max(m_{13}, m_{23}). \tag{21}$$

Furthermore

$$\begin{aligned} b_{12} &= \frac{\gamma_{313} - \gamma_{323}}{-\gamma_{212}} \\ &= \frac{-\gamma_{313} + \gamma_{323}}{\gamma_{212}} \\ &= \frac{-(\gamma_{232} - \gamma_{212})b_{13} + \gamma_{232}b_{23}}{\gamma_{212}} \\ &= \left(1 - \frac{\gamma_{232}}{\gamma_{212}}\right) b_{13} + \frac{\gamma_{232}}{\gamma_{212}} b_{23}. \end{aligned} \tag{22}$$

This is a weighted sum of the y-intercepts b_{13} and b_{23}, where the weights are positive and sum to one; thus, in addition to (21), we have the condition

$$b_{13} \leq b_{12} \leq b_{23} \tag{23}$$

since $b_{13} < b_{23}$ by (11) and (19).

If $m_{23} < 0$, then (21) immediately reduces to $m_{23} \leq m_{12} \leq m_{13}$ (by (18), we are considering a special case in which $m_{13} > 0$). This is illustrated in Fig. 3 for the slightly different situations $\chi_{13} < \chi_{23}$ and $\chi_{13} \geq \chi_{23}$. If, on the other hand, $m_{23} \geq 0$, then (21) and (23) together imply two possible situations, depending on whether $m_{23} < m_{13}$ or $m_{23} \geq m_{13}$. These possibilities are illustrated in Fig. 4.

One may of course ask what happens when $\gamma_{232} - \gamma_{212} = 0$ (ie, $\gamma_{232} = \gamma_{212}$, or $U_{1|2} = U_{3|2}$). In this case, both m_{13} and b_{13} are infinite. Furthermore

$$\begin{aligned} m_{23} &= \frac{-(\gamma_{131} - \gamma_{121})}{\gamma_{232}} \\ &= \frac{-\gamma_{131}}{\gamma_{232}} + \frac{\gamma_{121}}{\gamma_{212}} \\ &= \frac{-\gamma_{131}}{\gamma_{232}} + m_{12} \\ &\leq m_{12} \end{aligned} \tag{24}$$

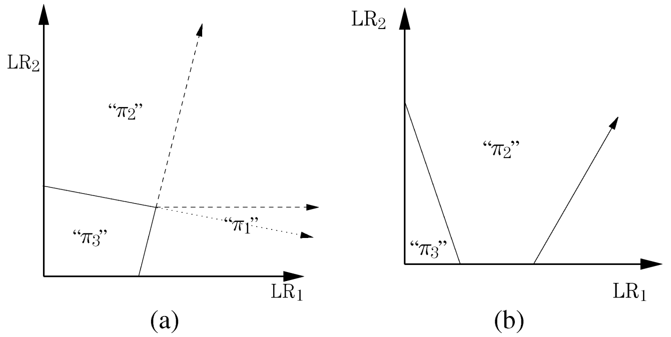


Fig. 3. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} < 0$ (implying $m_{13} > 0$ and $b_{13} < 0$) and $m_{23} < 0$. In (a), $\chi_{13} < \chi_{23}$, and the "1-vs-2" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $m_{12} > 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_1." In (b), $\chi_{13} \geq \chi_{23}$ and the "1-vs-2" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-3" and "2-vs-3" lines shown); observations above this line will be decided "π_2", and those below this line will be decided "π_1."

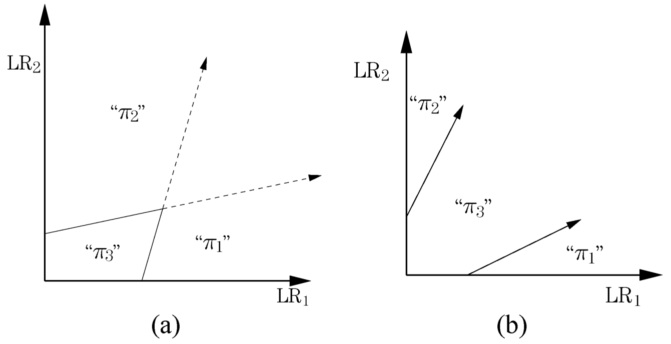


Fig. 4. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} < 0$ (implying $m_{13} > 0$ and $b_{13} < 0$) and $m_{23} \geq 0$. In (a), $m_{23} < m_{13}$, and the "1-vs-2" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2", and those below this line will be decided "π_1". In (b), $m_{23} \geq m_{13}$, and the "1-vs-2" line can lie anywhere between the "1-vs-3" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_3" regardless of the position of this line.

and

$$\begin{aligned} b_{12} &= \frac{\gamma_{323} - \gamma_{313}}{\gamma_{212}} \\ &= \frac{\gamma_{323}}{\gamma_{232}} + \frac{-\gamma_{313}}{\gamma_{212}} \\ &= b_{23} + \frac{-\gamma_{313}}{\gamma_{212}} \\ &\leq b_{23}. \end{aligned} \tag{25}$$

Together, (24) and (25) can be considered *either* a special case of the inequalities (15) and (17), if we take $m_{13} = -\infty$ and $b_{13} = +\infty$; *or* of the inequalities (21) and (23), if we take $m_{13} = +\infty$ and $b_{13} = -\infty$. This situation, for the slightly different cases $b_{12} < 0$ and $b_{12} \geq 0$, is illustrated in Fig. 5.

In this section, the possible values of the quantity $\gamma_{232} - \gamma_{212}$ were considered in order to determine properties of the ideal observer decision boundary lines. It may be argued that the choice of a parameter from the "1-vs-3" line, i.e., one of the three available lines, must be an arbitrary one. In fact, we may consider taking another parameter (or combination of parameters) from (6)–(8), and using it to determine conditions on the properties of the decision boundary lines as above. Given that all possible values of the quantity $\gamma_{232} - \gamma_{212}$ were considered, it is expected that no new conditions should be determinable (let alone conditions inconsistent with those already determined). In fact, this can readily be shown to be the case; however, due to the repetitive nature of the derivations involved, these are relegated to Appendices A and B.

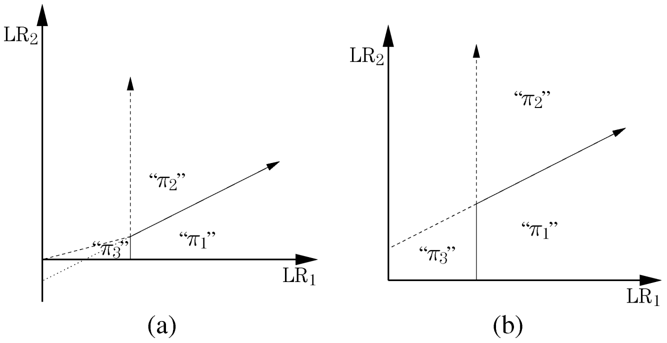


Fig. 5. Example ideal observer decision rules for the case $\gamma_{232} - \gamma_{212} = 0$ (implying $m_{13} = \mp\infty$ and $b_{13} = \pm\infty$). In (a), $b_{12} < 0$ and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the lower dashed and dotted lines is excluded because $b_{23} > 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $b_{12} \geq 0$ and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3."

IV. Discussion and Conclusion

The repetitive nature of the algebraic manipulations given in the preceding section and the Appendices should not be allowed to distract from the fundamental point being made: given the locations of two of the decision boundary lines, the location of the third is not completely arbitrary. That is, aside from the obvious [given (6)–(8)] constraint that the lines must share a common intersection point, it can also be shown that the slope of the third line is constrained by the slopes of the first two.

The significance of this result may be difficult to appreciate at first glance. It is perhaps best illustrated by comparison with the two-class classifier, for which the ROC operating point coordinates [e.g., the true-positive fraction (TPF) and false-positive fraction (FPF)] are determined by a single decision criterion γ, which is free to vary without restriction throughout its domain of definition. For the two-class ideal observer, in particular, an observation is decided "positive" (assigned to the class π_1) if LR$_1 > \gamma$, where γ can take on any nonnegative value. Furthermore, the FPF and TPF are related in a very simple way to the cdfs of $\mathbf{LR}_1$, and are thus monotonic in the decision criterion γ. For the three-class ideal observer, this straightforward relationship is lost; indeed, Figs. 2(b), 4(b), 7(b), 9(b), 12(b), and 14(b) show that for certain values of four of the five decision criteria γ_{iji}, the misclassification probabilities (ie, the ROC operating point coordinates) can be independent of the fifth decision criterion.

More succinctly, the relationship between the decision criteria and the misclassification probabilities is *not* one-to-one, as it is for the two-class ideal observer. A correct formulation of the misclassification probabilities as functions of the decision criteria—necessary for an explicit calculation of the ideal

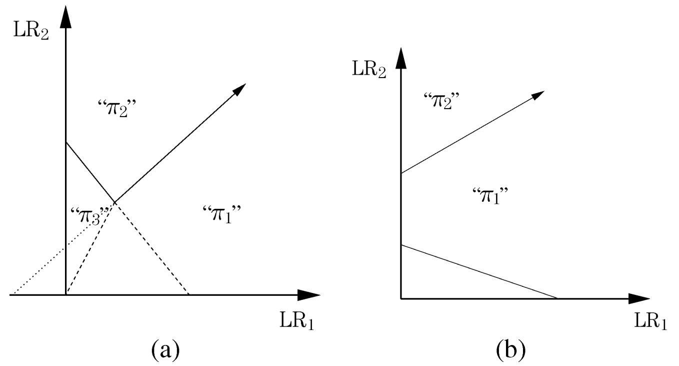


Fig. 6. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} > 0$ (implying $1/m_{23} < 0$ and $\chi_{23} > 0$) and $\chi_{12} < 0$. In (a), $b_{12} < b_{23}$, and the "1-vs-3" line can lie anywhere between the two dashed lines shown (the region between the left dashed and dotted lines is excluded because $\chi_{13} > 0$); observations in the unlabeled region to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3." In (b), $b_{12} \geq b_{23}$ and the "1-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "2-vs-3" lines shown); observations to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3."

observer's ROC hypersurface given the decision variable probability density functions—will require careful consideration of this issue. Although we have shown previously that the hypervolume under the ROC hypersurface is not a useful performance metric in general [19], it is still the case that the ROC hypersurface in terms of the set of misclassification probabilities (six in the three-class classification task) is a complete description of observer performance. We expect that a useful performance metric, assuming one exists, will be derived in some fashion from the ROC hypersurface. It is thus important to develop a complete understanding of the rather complicated relationships among the quantities involved, and we hope that this paper will prove of some use toward this goal.

Appendix A
Restrictions Determined by the Parameters of the "2-vs.-3" Line

Consider the quantity $\gamma_{131} - \gamma_{121}$ from (8). In particular, when $\gamma_{131} - \gamma_{121} > 0$ (ie, $\gamma_{131} > \gamma_{121}$, or $U_{2|1} > U_{3|1}$), we have

$$\frac{1}{m_{23}} = \frac{-\gamma_{232}}{\gamma_{131} - \gamma_{121}} < 0 \tag{26}$$

$$\chi_{23} = \frac{\gamma_{323}}{\gamma_{131} - \gamma_{121}} > 0. \tag{27}$$

Through reasoning similar to that of Section III, we also have

$$\frac{1}{m_{23}} \leq \frac{1}{m_{13}} \leq \frac{1}{m_{12}} \tag{28}$$

and

$$\min(\chi_{12}, \chi_{23}) \leq \chi_{13} \leq \max(\chi_{12}, \chi_{23}). \tag{29}$$

If $\chi_{12} < 0$, then (29) immediately reduces to $\chi_{12} \leq \chi_{13} \leq \chi_{23}$ (by (27), we are considering a special case in which $\chi_{23} > 0$). This is illustrated in Fig. 6 for the slightly different situations $b_{12} < b_{23}$ and $b_{12} \geq b_{23}$. If, on the other hand, $\chi_{12} \geq 0$, then (28) and (29) together imply two possible situations, depending on whether $\chi_{12} < \chi_{23}$ or $\chi_{12} \geq \chi_{23}$. These possibilities are illustrated in Fig. 7.

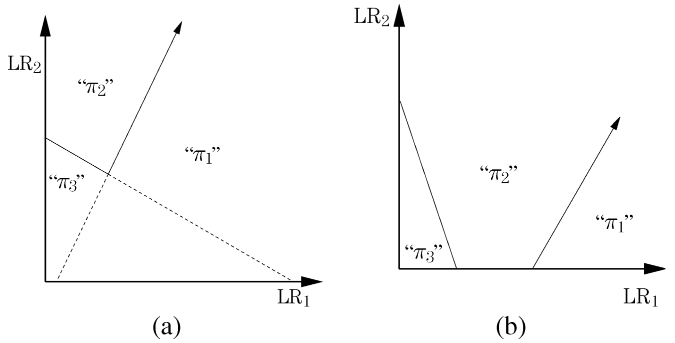


Fig. 7. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} > 0$ (implying $1/m_{23} < 0$ and $\chi_{23} > 0$) and $\chi_{12} \geq 0$. In (a), $\chi_{12} < \chi_{23}$, and the "1-vs-3" line can lie anywhere in the unlabeled region; observations to the left of this line will be decided "π_1," and those to the right of this line will be decided "π_3." In (b), $\chi_{12} \geq \chi_{23}$ and the "1-vs-3" line can lie anywhere between the "1-vs-2" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_2" regardless of the position of this line.

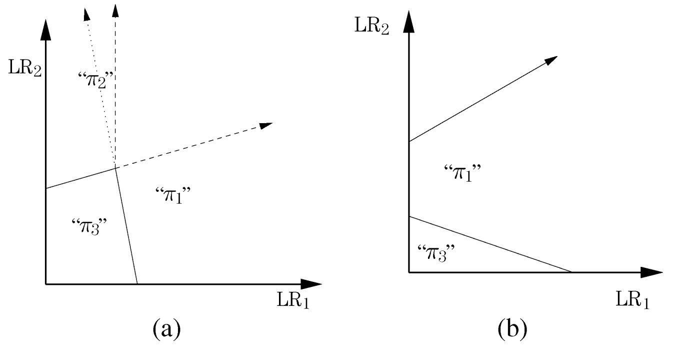


Fig. 8. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} < 0$ (implying $1/m_{23} > 0$ and $\chi_{23} < 0$) and $1/m_{13} < 0$. In (a), $b_{23} < b_{13}$, and the "1-vs-2" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $m_{12} > 0$ and, therefore, $1/m_{12} \geq 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_1." In (b), $b_{23} \geq b_{13}$ and the "1-vs-2" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-3" and "2-vs-3" lines shown); observations above this line will be decided "π_2", and those below this line will be decided "π_1."

If $\gamma_{131} - \gamma_{121} < 0$ (ie, $\gamma_{131} < \gamma_{121}$, or $U_{2|1} < U_{3|1}$), we have

$$\frac{1}{m_{23}} = \frac{-\gamma_{232}}{\gamma_{131} - \gamma_{121}} > 0 \tag{30}$$

$$\chi_{23} = \frac{\gamma_{323}}{\gamma_{131} - \gamma_{121}} < 0. \tag{31}$$

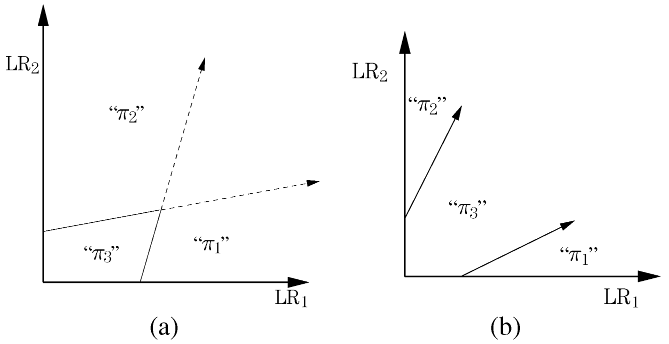
One can also show

$$\min\left(\frac{1}{m_{13}}, \frac{1}{m_{23}}\right) \leq \frac{1}{m_{12}} \leq \max\left(\frac{1}{m_{13}}, \frac{1}{m_{23}}\right) \tag{32}$$

and

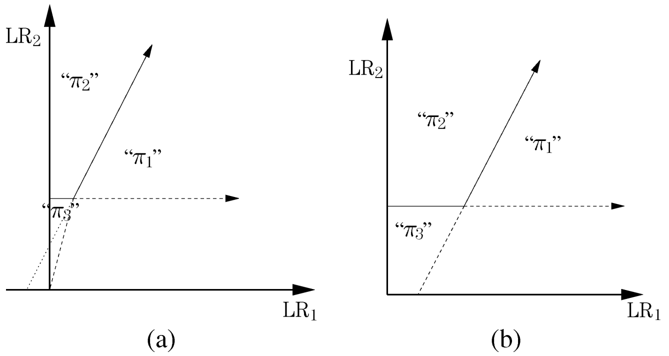
$$\chi_{23} \leq \chi_{12} \leq \chi_{13}. \tag{33}$$

If $1/m_{13} < 0$, then (32) immediately reduces to $1/m_{13} \leq 1/m_{12} \leq 1/m_{23}$ (by (30), we are considering a special case in which $1/m_{23} > 0$). This is illustrated in Fig. 8 for the slightly different situations $b_{23} < b_{13}$ and $b_{23} \geq b_{13}$. If, on the other



(a) (b)

Fig. 9. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} < 0$ (implying $1/m_{23} > 0$ and $\chi_{23} < 0$) and $1/m_{13} \geq 0$. In (a), $1/m_{13} < 1/m_{23}$, and the "1-vs-2" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_1." In (b), $1/m_{13} \geq 1/m_{23}$ and the "1-vs-2" line can lie anywhere between the "1-vs-3" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_3" regardless of the position of this line.



(a) (b)

Fig. 10. Example ideal observer decision rules for the case $\gamma_{131} - \gamma_{121} = 0$ (implying $1/m_{23} = \mp\infty$ and $\chi_{23} = \pm\infty$). In (a), $\chi_{12} < 0$, and the "1-vs-3" line can lie anywhere between the two dashed lines shown (the region between the leftmost dashed and dotted lines is excluded because $\chi_{13} > 0$); observations in the unlabeled region to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3." In (b), $\chi_{12} \geq 0$ and the "1-vs-3" line can lie anywhere in the unlabeled region; observations to the right of this line will be decided "π_1," and those to the left of this line will be decided "π_3."

hand, $1/m_{13} \geq 0$, then (32) and (33) together imply two possible situations, depending on whether $1/m_{13} < 1/m_{23}$ or $1/m_{13} \geq 1/m_{23}$. These possibilities are illustrated in Fig. 9.

Finally, we consider the case $\gamma_{131} - \gamma_{121} = 0$ ($\gamma_{131} = \gamma_{121}$ or $U_{2|1} = U_{3|1}$), in which both $1/m_{23}$ and χ_{23} are infinite. We now have

$$\frac{1}{m_{13}} \leq \frac{1}{m_{12}} \tag{34}$$

and

$$\chi_{12} \leq \chi_{13}. \tag{35}$$

Together, (34) and (35) can be considered *either* a special case of the inequalities (28) and (29), if we take $1/m_{23} = -\infty$ and $\chi_{23} = +\infty$; *or* of the inequalities (32) and (33), if we take $1/m_{23} = +\infty$ and $\chi_{23} = -\infty$. This situation, for the slightly different cases $\chi_{12} < 0$ and $\chi_{12} \geq 0$, is illustrated in Fig. 10.

Notice that every figure in this appendix has one or more corresponding figures in Section III (depending on the possible values of the undetermined decision boundary parameter being illustrated in that figure). Specifically

Fig. 6(a)	⇒	Figs. 2(a), 3(a), 5(b)
Fig. 6(b)	⇒	Fig. 2(b)
Fig. 7(a)	⇒	Figs. 1(a), 3(a), 5(a)
Fig. 7(b)	⇒	Figs. 1(b), 3(b), 5(a)
Fig. 8(a)	⇒	Figs. 1(a), 2(a)
Fig. 8(b)	⇒	Fig. 2(b)
Fig. 9(a)	⇒	Figs. 4(a), 5(a), 5(b)
Fig. 9(b)	⇒	Fig. 4(b)
Fig. 10(a)	⇒	Figs. 2(a), 4(a), 5(b), 2(b)
Fig. 10(b)	⇒	Figs. 1(a), 4(a), 5(a).

That is, none of the conditions derived in this section are inconsistent with those derived Section III. More importantly, note the symmetry between the corresponding equations and figures in Section III and this appendix, if one "swaps" the labels of classes π_1 and π_2, and additionally replaces m_{ij} with $1/m_{i'j'}$, χ_{ij} with $b_{i'j'}$, and b_{ij} with $\chi_{i'j'}$ ($i' = 1$ if $i = 2$, 2 if $i = 1$, and 3 if $i = 3$; similarly for j). Intuitively, if one "flips" the figures in one section about the $y = x$ line, one obtains the figures in the other section.

Appendix B
Restrictions Determined by the Parameters of the "1-vs.-2" Line

In this appendix, we consider the possible values of the quantity $\gamma_{313} - \gamma_{323}$. As in the preceding Appendix, we expect to obtain no conditions inconsistent with those already derived.

When $\gamma_{313} - \gamma_{323} > 0$ (ie, $\gamma_{313} > \gamma_{323}$, or $U_{2|3} > U_{1|3}$), we have

$$\frac{1}{b_{12}} = \frac{-\gamma_{212}}{\gamma_{313} - \gamma_{323}} < 0 \tag{36}$$

$$\frac{1}{\chi_{12}} = \frac{\gamma_{121}}{\gamma_{313} - \gamma_{323}} > 0. \tag{37}$$

Through reasoning similar to that of Section III, we also have

$$\frac{1}{b_{12}} \leq \frac{1}{b_{13}} \leq \frac{1}{b_{23}} \tag{38}$$

and

$$\min\left(\frac{1}{\chi_{23}}, \frac{1}{\chi_{12}}\right) \leq \frac{1}{\chi_{13}} \leq \max\left(\frac{1}{\chi_{23}}, \frac{1}{\chi_{12}}\right). \tag{39}$$

If $1/\chi_{23} \leq 0$, then (39) immediately reduces to $1/\chi_{23} \leq 1/\chi_{13} \leq 1/\chi_{12}$ (by (37), we are considering a special case in which $1/\chi_{12} > 0$). This is illustrated in Fig. 11 for the slightly different situations $m_{23} < m_{12}$ and $m_{23} \geq m_{12}$. If, on the other hand, $1/\chi_{23} > 0$, then (38) and (39) together imply two possible situations, depending on whether $1/\chi_{23} < 1/\chi_{12}$ or $1/\chi_{23} \geq 1/\chi_{12}$. These possibilities are illustrated in Fig. 12.

If $\gamma_{313} - \gamma_{323} < 0$ (ie, $\gamma_{313} < \gamma_{323}$, or $U_{2|3} < U_{1|3}$), we have

$$\frac{1}{b_{12}} = \frac{-\gamma_{212}}{\gamma_{313} - \gamma_{323}} > 0 \tag{40}$$

$$\frac{1}{\chi_{12}} = \frac{\gamma_{121}}{\gamma_{313} - \gamma_{323}} < 0. \tag{41}$$

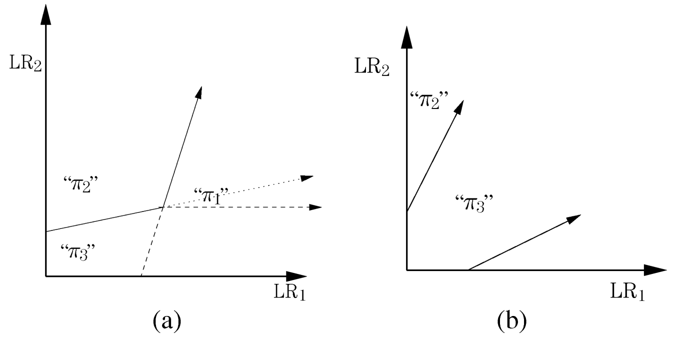


Fig. 11. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} > 0$ (implying $1/b_{12} < 0$ and $1/\chi_{12} > 0$) and $1/\chi_{23} \leq 0$. In (a), $m_{23} < m_{12}$, and the "1-vs-3" line can lie anywhere between the two dashed lines shown (the region between the horizontal dashed and dotted lines is excluded because $\chi_{13} > 0$ and, therefore, $1/\chi_{13} \geq 0$); observations in the unlabeled region to the left of this line will be decided "π_3", and those to the right of line will be decided "π_1." In (b), $m_{23} \geq m_{12}$, and the "1-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "2-vs-3" lines shown); observations to the left of this line will be decided "π_3," and those to the right of this line will be decided "π_1."

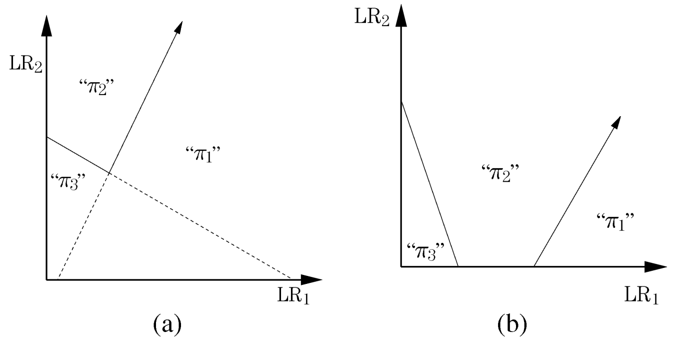


Fig. 12. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} > 0$ (implying $1/b_{12} < 0$ and $1/\chi_{12} > 0$) and $1/\chi_{23} > 0$. In (a), $1/\chi_{23} < 1/\chi_{12}$ and the "1-vs-3" line can lie anywhere in the unlabeled region; observations to the left of this line will be decided "π_3," and those to the right of this line will be decided "π_1." In (b), $1/\chi_{23} \geq 1/\chi_{12}$, and the "1-vs-3" line can lie anywhere between the "1-vs-2" and "2-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_2" regardless of the position of this line.

One can also show

$$\min\left(\frac{1}{b_{13}}, \frac{1}{b_{12}}\right) \leq \frac{1}{b_{23}} \leq \max\left(\frac{1}{b_{13}}, \frac{1}{b_{12}}\right) \qquad (42)$$

and

$$\frac{1}{\chi_{12}} \leq \frac{1}{\chi_{23}} \leq \frac{1}{\chi_{13}}. \qquad (43)$$

If $1/b_{13} \leq 0$, then (42) immediately reduces to $1/b_{13} \leq 1/b_{23} \leq 1/b_{12}$ (by (40), we are considering a special case in which $1/b_{12} > 0$). This is illustrated in Fig. 13 for the slightly different situations $m_{12} < m_{13}$ and $m_{12} \geq m_{13}$. If, on the other hand, $1/b_{13} > 0$, then (42) and (43) together imply two possible situations, depending on whether $1/b_{13} < 1/b_{12}$ or $1/b_{13} \geq 1/b_{12}$. These possibilities are illustrated in Fig. 14.

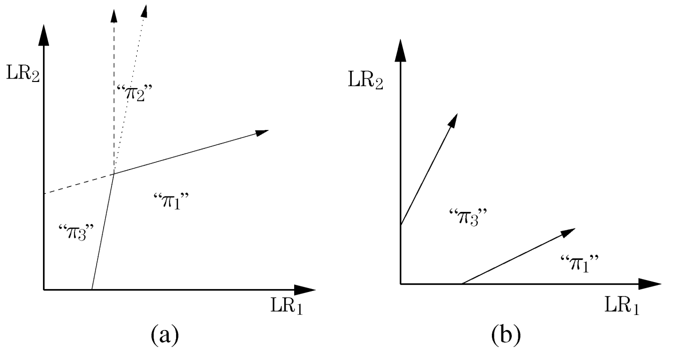


Fig. 13. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} < 0$ (implying $1/b_{12} > 0$ and $1/\chi_{12} < 0$) and $1/b_{13} \leq 0$. In (a), $m_{12} < m_{13}$, and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $b_{23} > 0$, and therefore $1/b_{23} \geq 0$); observations in the unlabeled region above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $m_{12} \geq m_{13}$, and the "2-vs-3" line can lie anywhere in the unlabeled region (provided it shares the intersection point of the "1-vs-2" and "1-vs-3" lines shown); observations above this line will be decided "π_2," and those below this line will be decided "π_3."

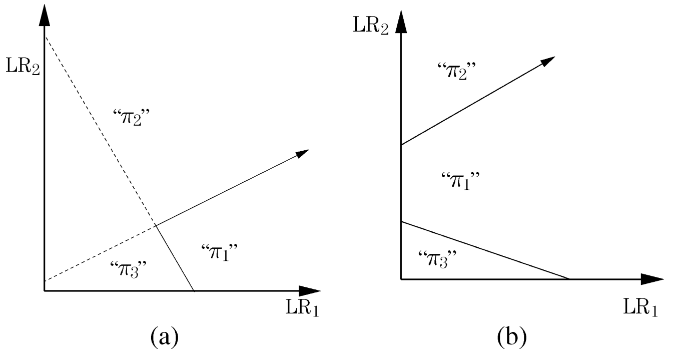


Fig. 14. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} < 0$ (implying $1/b_{12} > 0$ and $1/\chi_{12} < 0$) and $1/b_{13} > 0$. In (a), $1/b_{13} < 1/b_{12}$, and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3". In (b), $1/b_{13} \geq 1/b_{12}$, and the "2-vs-3" line can lie anywhere between the "1-vs-2" and "1-vs-3" lines (provided it shares their intersection point); note that observations in this region will be decided "π_1" regardless of the position of this line.

Finally, we consider the case $\gamma_{323} - \gamma_{313} = 0$ (ie, $\gamma_{313} = \gamma_{323}$, or $U_{2|3} = U_{1|3}$), in which both $1/b_{12}$ and $1/\chi_{12}$ are infinite. We now have

$$\frac{1}{b_{13}} \leq \frac{1}{b_{23}} \qquad (44)$$

and

$$\frac{1}{\chi_{23}} \leq \frac{1}{\chi_{13}}. \qquad (45)$$

Together, (44) and (45) can be considered *either* a special case of the inequalities (38) and (39), if we take $1/b_{12} = -\infty$ and $1/\chi_{12} = +\infty$; *or* of the inequalities (42) and (43), if we take $1/b_{12} = +\infty$ and $1/\chi_{12} = -\infty$. This situation, for the slightly different cases $1/b_{13} \leq 0$ and $1/b_{13} > 0$, is illustrated in Fig. 15.

Notice that every figure in this appendix has one or more corresponding figures in Section III (depending on the possible

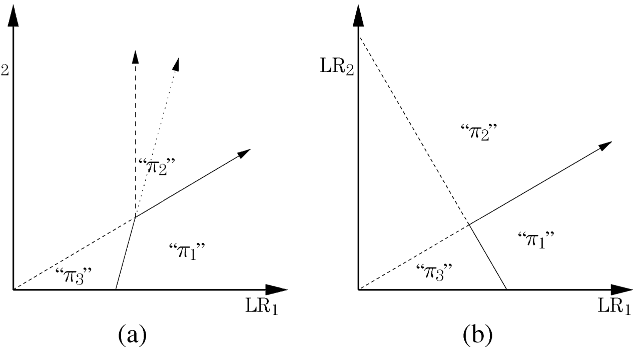


Fig. 15. Example ideal observer decision rules for the case $\gamma_{313} - \gamma_{323} = 0$ (implying $1/b_{12} = \mp\infty$ and $1/\chi_{12} = \pm\infty$). In (a), $1/b_{13} \leq 0$, and the "2-vs-3" line can lie anywhere between the two dashed lines shown (the region between the vertical dashed and dotted lines is excluded because $1/b_{23} \geq 0$); observations in the unlabeled region to above this line will be decided "π_2," and those below this line will be decided "π_3." In (b), $1/b_{13} > 0$, and the "2-vs-3" line can lie anywhere in the unlabeled region; observations above this line will be decided "π_2," and those below this line will be decided "π_3."

values of the undetermined decision boundary parameter being illustrated in that figure). Specifically

$$\begin{array}{lcl}
\text{Fig. } 11(\text{a}) & \Rightarrow & \text{Figs. } 1(\text{a}),\ 4(\text{a}),\ 5(\text{a}) \\
\text{Fig. } 11(\text{b}) & \Rightarrow & \text{Fig. } 4(\text{b}) \\
\text{Fig. } 12(\text{a}) & \Rightarrow & \text{Figs. } 1(\text{a}),\ 3(\text{a}),\ 5(\text{a}) \\
\text{Fig. } 12(\text{b}) & \Rightarrow & \text{Figs. } 1(\text{b}),\ 3(\text{b}),\ 5(\text{a}) \\
\text{Fig. } 13(\text{a}) & \Rightarrow & \text{Figs. } 3(\text{a}),\ 4(\text{a}),\ 5(\text{b}) \\
\text{Fig. } 13(\text{b}) & \Rightarrow & \text{Fig. } 4(\text{b}) \\
\text{Fig. } 14(\text{a}) & \Rightarrow & \text{Fig. } 2(\text{a}) \\
\text{Fig. } 14(\text{b}) & \Rightarrow & \text{Fig. } 2(\text{b}) \\
\text{Fig. } 15(\text{a}) & \Rightarrow & \text{Figs. } 3(\text{a}),\ 4(\text{a}),\ 5(\text{b}) \\
\text{Fig. } 15(\text{b}) & \Rightarrow & \text{Figs. } 2(\text{a}),\ 3(\text{a}),\ 4(\text{b}).
\end{array}$$

That is, none of the conditions derived in this appendix are inconsistent with those derived in Section III or Appendix A. More importantly, note the symmetry between the corresponding equations and figures in Sections III and this appendix, if one "swaps" the labels of classes π_2 and π_3, and additionally replaces m_{ij} with $1/\chi_{i'j'}$, χ_{ij} with $1/m_{i'j'}$, and b_{ij} with $1/b_{i'j'}$ ($i' = 1$ if $i = 1$, 2 if $i = 3$, and 3 if $i = 2$; similarly for j).

Acknowledgment

The authors thank the associate editor and anonymous reviewers for their suggestions to substantially improve the content and structure of this manuscript.

References

[1] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic, 1975.

[2] C. E. Metz, "Basic principles of ROC analysis," *Sem. Nucl. Med.*, vol. VIII, no. 4, pp. 283–298, 1978.

[3] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*. New York: Wiley, 1968.

[4] B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math Psych.*, vol. 40, pp. 253–269, 1996.

[5] ——, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psych.*, vol. 42, pp. 5–31, 1998.

[6] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, pp. 78–89, 1999.

[7] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study," *Proc. SPIE Medical Imaging 2003: Image Processing*, vol. 5032, pp. 567–578, 2003.

[8] U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.*, vol. 2, pp. 1–9, 1995.

[9] F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images," *Med. Phys.*, vol. 18, pp. 955–963, 1991.

[10] F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Invest. Radiol.*, vol. 28, pp. 473–481, 1993.

[11] F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: automated alignment of breast images and its effect on bilateral-subtraction technique," *Med. Phys.*, vol. 21, pp. 445–452, 1994.

[12] M. A. Kupinski, "Computerized pattern classification in medical imaging," Ph.D. thesis, The Univ. Chicago, Chicago, IL, 2000.

[13] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[14] Z. Huo, M. L. Giger, and C. E. Metz, "Effect of dominant features on neural network performance in the classification of mammographic lesions," *Phys. Med. Biol.*, vol. 44, pp. 2579–2595, 1999.

[15] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, "Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness," *Acad. Radiol.*, vol. 7, pp. 1077–1084, 2000.

[16] Z. Huo, M. L. Giger, and C. J. Vyborny, "Computerized analysis of multiple-mammographic views: potential usefulness of special view mammograms in computer-aided diagnosis," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1285–1292, Dec. 2001.

[17] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: effectiveness of computer-aided diagnosis—observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560–568, 2002.

[18] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 891–895, Jul. 2004.

[19] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks," *IEEE Trans. Med. Imag.*, vol. 24, no. 3, pp. 293–299, Mar. 2005.

[20] A. Srinivasan, "Note on the Location of Optimal Classifiers in n-Dimensional ROC Space," Oxford Univ. Computing Lab., Oxford, U.K., Tech. Rep. PRG-TR-2-99, 1999.

[21] C. Ferri, J. Hernández-Orallo, and M. A. Salido, "Volume Under the roc Surface for Multi-Class Problems: Exact Computation and Evaluation of Approximations," Dep. Sistemes Informàtics i Computació, Univ. Politècnica de València, València, Spain, Tec. Rep. 2003.

[22] C. E. Metz, "The optimal decision variable," Dept. Radiol., Univ. Chicago, unpublished lecture notes for the course Mathematics for Medical Physicists. 2000.

H Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities

Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities

Darrin C. Edwards* and Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

We have shown in previous work that an ideal observer in a classification task with N classes achieves the optimal receiver operating characteristic (ROC) hypersurface in a Neyman-Pearson sense. That is, the hypersurface obtained by taking one of the ideal observer's misclassification probabilities as a function of the other $N^2 - N - 1$ misclassification probabilities is never above the corresponding hypersurface obtained by any other observer. Due to the inherent complexity of evaluating observer performance in an N-class classification task with $N > 2$, some researchers have suggested a generally incomplete but more tractable evaluation in terms of a hypersurface plotting only the N "sensitivities" (the probabilities of correctly classifying observations in the various classes). An N-class observer generally has up to $N^2 - N - 1$ degrees of freedom, so a given sensitivity will still vary when the other $N - 1$ are held fixed; a well-defined hypersurface can be constructed by considering only the maximum possible value of one sensitivity for each achievable value of the other $N - 1$. We show that optimal performance in terms of this generally incomplete performance descriptor, in a Neyman-Pearson sense, is still achieved by the N-class ideal observer. That is, the hypersurface obtained by taking the maximal value of one of the ideal observer's correct classification probabilities as a function of the other $N - 1$ is never below the corresponding hypersurface obtained by any other observer.

Keywords: ROC analysis, three-class classification, ideal observer decision rules

1. INTRODUCTION

We are attempting to extend the well-known observer performance evaluation methodology of receiver operating characteristic (ROC) analysis[1, 2] to classification tasks with three classes. This could conceivably be of benefit, for example, in a medical decision-making task in which a region of a patient image must be characterized as containing a malignant lesion, a benign lesion, or only normal tissue.[3]

Unfortunately, a fully general but tractable extension of ROC analysis has yet to be developed. It is known that the performance of an observer in a classification task with N classes ($N \geq 2$) can be completely described by a set of $N^2 - N$ conditional error probabilities,[4, 5] and that the performance of the ideal observer (that which minimizes Bayes risk[4]) is completely characterized by an ROC hypersurface in which these conditional error probabilities depend on a set of $N^2 - N - 1$ decision criteria.[5] Although analytic expressions for the ideal observer's conditional error probabilities given reasonable models for the underlying observational date have been worked out in the two-class case,[6] this has not yet been accomplished in a fully general manner for tasks with three or more classes. Furthermore, we have shown that an obvious generalization of the area under the ROC curve (AUC) does not in fact yield a useful performance metric in tasks with three or more classes.[7] More recently, we showed that complicated constraining relationships exist among the decision criteria themselves for the ideal observer.[8] These constraining relationships appear to imply that it is highly unlikely that analytical expressions for the conditional error probabilities in terms of the decision criteria can be developed which are as simple to interpret as those for the two-class task.[6]

Despite the difficulties just described, the potential benefits to be gained from a practical performance evaluation methodology for classification tasks with three classes have motivated a number of research groups to propose such methods. These practical methods reduce the number of degrees of freedom required to describe the observer's performance, either by implicitly leaving the remaining degrees of freedom out of the analysis, or

*Correspondence: E-mail: d-edwards@uchicago.edu; Telephone: 773 834 5094; Fax: 773 702 0371

by explicitly imposing restrictions on the form of the observer's decision rule or on the set of decision criteria used by the observer.

Scurfield evaluated an observer which used a specified decision rule with only two degrees of freedom (as opposed to the five decision criteria used by the general three-class ideal observer) by plotting a set of six (two-dimensional) surfaces in three-dimensional ROC spaces.[9] Mossman proposed plotting the surface formed only from the set of three "sensitivities" (conditional probabilities of correctly classifying observations) for an observer with two degrees of freedom, and applied this method to an observer with a specified decision rule.[10] Chan *et al.* began with an ideal observer model, and reduced the number of decision criteria from five to two by imposing explicit assumptions on the observer's decision utilities; the observer's performance was then plotted as a surface in a three-dimensional ROC space, the axes of which are the probabilities of deciding an observation to be malignant conditional on each of the three actual class memberships.[11] He *et al.* investigated an ideal observer model in which the decision rule is restricted to a form similar to that proposed by Scurfield; the nature of the restrictions is such that performance evaluation in terms of only the three sensitivities provides a complete description of this observer's performance.[12]

A common theme among these remarkably diverse methods is the idea of an "ROC surface," *i.e.*, a surface with two degrees of freedom in a three-dimensional ROC space. An appealing feature of such a construct is its visualizability: it can be plotted as readily as any elevation map, for example, in stark contrast to the fully general three-class classification task involving a hypersurface with five degrees of freedom in a six-dimensional ROC space as mentioned above. While it is true that not all of the proposed methods described in the preceding paragraph involve a "sensitivity" ROC surface, the general division of an N-class observer's conditional decision probabilities into a set of N sensitivities and a set of $N^2 - N$ misclassification rates[5] makes this particular construct a natural candidate for further analysis.

On the other hand, it can be argued that measurement of performance in terms of only N conditional classification rates must be an incomplete description of observer performance in a classification task with more than two classes, which requires $N^2 - N$ such classification rates as stated above. Acknowledging this incompleteness, we would like to ask whether there is any sense in which such an incomplete performance metric is at least well-defined. In particular, is there any observer decision rule, dependent on only $N - 1$ (rather than $N^2 - N - 1$) decision criteria, for which the observer's sensitivity ROC hypersurface is always above the corresponding hypersurface obtained for any other observer? If so, what form does this decision rule take?

In the next section, we show that the three-class observer which optimizes performance only in terms of the sensitivity surface is in fact the three-class ideal observer, with its decision utilities constrained in a particular way (reducing its degrees of freedom from five to two as necessary). Additionally, the form of the constraints on the ideal observer's behavior are identical to those considered by He *et al.*.[12] In Sec. 3, we extend this result to the general case of an N-class observer, showing that the observer which attains the optimal sensitivity hypersurface is a restricted form of the N-class ideal observer, and in particular a straightforward generalization of the three-class observer considered by He *et al.*[12] to N classes. Our conclusions are stated in Sec. 4.

2. THREE-CLASS OBSERVERS

We have shown[5] that the N-class ideal observer — that observer which minimizes Bayes risk — also achieves optimal performance in an ROC sense, by virtue of satisfying the Neyman-Pearson criterion. This was the same argument used by Van Trees[4] to show that the two-class ideal observer achieves the optimal ROC curve for a given two-class classification task. This technique of satisfying the Neyman-Pearson criterion, essentially an application of an integral form of the method of Lagrange multipliers,[13] is straightforward (conceptually, if not notationally) and flexible, and we apply it in this section to answer the question of what observer optimizes performance in terms of only the three observer sensitivities.

We denote by P_{ij} the conditional probability of a given observer deciding an observation is drawn from the ith class, conditional on it actually being drawn from the jth class. Thus, the three sensitivities are P_{11}, P_{22}, and P_{33}. Decisions are assumed to be made based on statistically variable observational data; in particular,

$$P_{ij} \equiv \int_{Z_i} p(\vec{x}|\pi_j)\, d^m\vec{x}, \tag{1}$$

where Z_i is the region for which observations $\vec{\mathbf{x}}$ (of dimension m) are decided to belong to the class labeled π_i $(1 \leq i \leq 3)$.

Without loss of generality, we seek to maximize P_{33} subject to the constraints $P_{11} = \alpha_{11}$ and $P_{22} = \alpha_{22}$ where $0 \leq \alpha_{11} \leq 1$ and $0 \leq \alpha_{22} \leq 1$. We define the function

$$F \equiv P_{33} + \lambda_{11}(P_{11} - \alpha_{11}), + \lambda_{22}(P_{22} - \alpha_{22}) \tag{2}$$

where λ_{11} and λ_{22} are the so-called Lagrange multipliers. Note that if we can find a decision rule (a partitioning of the domain of $\vec{\mathbf{x}}$ into Z_1, Z_2, and Z_3) that maximizes F for arbitrary values of λ_{11} and λ_{22}, then this will be equivalent to maximizing P_{33} at the point at which the constrain equations are satisfied (*i.e.*, at the point $P_{11} = \alpha_{11}, P_{22} = \alpha_{22}$).

We first rewrite F by applying rules for conditional probabilities:

$$\begin{aligned} F &= -\lambda_{11}\alpha_{11} - \lambda_{22}\alpha_{22} + (1 - P_{13} - P_{23}) + \lambda_{11}(1 - P_{21} - P_{31}) + \lambda_{22}(1 - P_{12} - P_{32}) \\ &= 1 + \lambda_{11}(1 - \alpha_{11}) + \lambda_{22}(1 - \alpha_{22}) - \{\lambda_{22}P_{12} + P_{13} + \lambda_{11}P_{21} + P_{23} + \lambda_{11}P_{31} + \lambda_{22}P_{32}\} \\ &= 1 + \lambda_{11}(1 - \alpha_{11}) + \lambda_{22}(1 - \alpha_{22}) - \left\{ \int_{Z_1} \lambda_{22} p(\vec{x}|\pi_2) + p(\vec{x}|\pi_3)\, d^m\vec{x} \right. \\ &\quad \left. + \int_{Z_2} \lambda_{11} p(\vec{x}|\pi_1) + p(\vec{x}|\pi_3)\, d^m\vec{x} + \int_{Z_3} \lambda_{11} p(\vec{x}|\pi_1) + \lambda_{22} p(\vec{x}|\pi_2)\, d^m\vec{x} \right\}. \end{aligned} \tag{3}$$

For a given set of values of the parameters λ_{11} and λ_{22}, F is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in Eq. 3) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{array}{llllll} \text{decide} & \pi_1 & \text{iff} & \lambda_{22} p(\vec{x}|\pi_2) < \lambda_{11} p(\vec{x}|\pi_1) & \text{and} & p(\vec{x}|\pi_3) < \lambda_{11} p(\vec{x}|\pi_1) \end{array} \tag{4}$$

$$\begin{array}{llllll} \text{decide} & \pi_2 & \text{iff} & \lambda_{11} p(\vec{x}|\pi_1) \leq \lambda_{22} p(\vec{x}|\pi_2) & \text{and} & p(\vec{x}|\pi_3) < \lambda_{22} p(\vec{x}|\pi_2) \end{array} \tag{5}$$

$$\begin{array}{llllll} \text{decide} & \pi_3 & \text{iff} & \lambda_{11} p(\vec{x}|\pi_1) \leq p(\vec{x}|\pi_3) & \text{and} & \lambda_{22} p(\vec{x}|\pi_2) \leq p(\vec{x}|\pi_3). \end{array} \tag{6}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\begin{array}{llllll} \text{decide} & \pi_1 & \text{iff} & \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 > 0 & \text{and} & \lambda_{11}\text{LR}_1 > 1 \end{array} \tag{7}$$

$$\begin{array}{llllll} \text{decide} & \pi_2 & \text{iff} & \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 \leq 0 & \text{and} & \lambda_{22}\text{LR}_2 > 1 \end{array} \tag{8}$$

$$\begin{array}{llllll} \text{decide} & \pi_3 & \text{iff} & \lambda_{11}\text{LR}_1 \leq 1 & \text{and} & \lambda_{22}\text{LR}_2 \leq 1, \end{array} \tag{9}$$

where $\text{LR}_i \equiv p(\vec{x}|\pi_i)/p(\vec{x}|\pi_3)$ are the likelihood ratio decision variables used by the ideal observer.[4,5] The decision boundary lines which partition the $(\text{LR}_1, \text{LR}_2)$ decision plane into the regions Z_1, Z_2, and Z_3 are thus

$$\lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 = 0 \tag{10}$$

$$\lambda_{11}\text{LR}_1 = 1 \tag{11}$$

$$\lambda_{22}\text{LR}_2 = 1. \tag{12}$$

Note that Eq. 12 is just the difference between Eqs. 10 and 11. If we require λ_{11} and λ_{22} to be positive, the decision rule is an ideal observer decision rule.[5] Since neither the decision variables nor the form of the decision rule depend on the particular choices of α_{11} and α_{22}, we can conclude that the three-class sensitivity ROC surface, obtained by allowing λ_{11} and λ_{22} to take on all possible positive values, is optimal for the observer defined in Eqs. 10–12, in the sense that no other observer can achieve a higher sensitivity surface (*i.e.*, a surface with a greater value of P_{33} at a given value of (P_{11}, P_{22})). The optimal observer for this performance metric is seen to be the three-class ideal observer, with its decision criteria constrained so that the line separating classes π_1 and π_3 is vertical, the line separating classes π_2 and π_3 is horizontal, and the line separating classes π_1 and

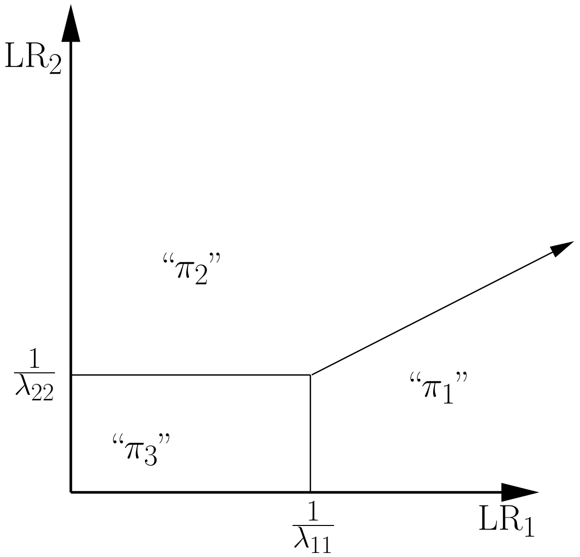


Figure 1. The decision rule which is found to be optimal in the sense of maximizing the ROC surface composed of only the observer sensitivities. The decision variables are the likelihood ratios used by the general three-class ideal observer, and the number of decision criteria is reduced from five (for the general three-class ideal observer) to two.

π_2 passes through the origin with slope $\lambda_{11}/\lambda_{22}$ (and thus intersects the other two lines as required). Note that the number of free decision criteria has been reduced from five (for the general three-class ideal observer) to two (as expected for a surface in a three-dimensional ROC space).

This decision rule is shown in Fig. 1. It is interesting to note that this observer is identical to the special case of the ideal observer evaluated by He *et al.*,[12] which we have shown[14, 15] to be a special case of the decision rule proposed by Scurfield.[9]

3. *N*-CLASS OBSERVERS

The results of the preceding section can be generalized to tasks with N classes for any $N > 2$. We now have a set of N^2 conditional classification probabilities P_{ij}, with N sensitivities P_{ii}. Equation 1 remains unchanged, except that there are of course now N regions Z_i into which the domain of $\vec{x}$ is partitioned (*i.e.*, classes into which the observations are classified), and the observations are drawn from N distributions of the form $p(\vec{x}|\pi_j)$.

Without loss of generality, we seek to maximize P_{NN} subject to the constraints $P_{ii} = \alpha_{ii}$ for $1 \leq i \leq N-1$, where $0 \leq \alpha_{ii} \leq 1$. We define the function

$$F \equiv P_{NN} + \sum_{i=1}^{N-1} \lambda_{ii}(P_{ii} - \alpha_{ii}), \tag{13}$$

where the λ_{ii} are the Lagrange multipliers. Note that if we can find a decision rule (a partitioning of the domain of $\vec{\mathbf{x}}$ into Z_i $\{1 \leq i \leq N\}$) that maximizes F for arbitrary values of the λ_{ii}, then this will be equivalent to maximizing P_{NN} at the point at which the constrain equations are satisfied (*i.e.*, at the point $P_{ii} = \alpha_{ii}$ $\{1 \leq i \leq N-1\}$).

As in the preceding section, we rewrite F by applying rules for conditional probabilities to obtain:

$$F = -\sum_{i=1}^{N-1} \lambda_{ii}\alpha_{ii} + \left(1 - \sum_{i=1}^{N-1} P_{iN}\right) + \sum_{i=1}^{N-1} \lambda_{ii} \left(1 - \sum_{\substack{j=1 \\ j\neq i}}^{N} P_{ji}\right)$$

$$
\begin{aligned}
&= 1+\sum_{i=1}^{N-1}\lambda_{ii}(1-\alpha_{ii})-\left\{\left[\sum_{i=1}^{N-1}\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\lambda_{jj}P_{ij}\right)+P_{iN}\right]+\left[\sum_{i=1}^{N-1}\lambda_{ii}P_{Ni}\right]\right\}\\
&= 1+\sum_{i=2}^{N}\lambda_{ii}(1-\alpha_{ii})\\
&\quad -\left\{\sum_{i=1}^{N-1}\int_{Z_i}\left[\sum_{\substack{j=1\\ j\neq i}}^{N}\lambda_{jj}p(\vec{x}|\pi_j)\right]+p(\vec{x}|\pi_N)\,d^m\vec{x}+\int_{Z_N}\sum_{i=1}^{N-1}\lambda_{ii}p(\vec{x}|\pi_i)\,d^m\vec{x}\right\}. \qquad (14)
\end{aligned}
$$

For a given set of values of the parameters λ_{ii} $\{1 \leq i \leq N-1\}$, F is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning choosing the regions Z_i such that a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in Eq. 14) is minimized. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$
\begin{aligned}
\text{decide}\quad \pi_i\{i<N\}\quad \text{iff}\quad & \lambda_{jj}p(\vec{x}|\pi_j) < \lambda_{ii}p(\vec{x}|\pi_i) \qquad \{i<j<N\}\\
\text{and}\quad & p(\vec{x}|\pi_N) < \lambda_{ii}p(\vec{x}|\pi_i)\\
\text{and}\quad & \lambda_{jj}p(\vec{x}|\pi_j) \leq \lambda_{ii}p(\vec{x}|\pi_i) \qquad \{j<i<N\} \qquad (15)\\
\text{decide}\quad \pi_N\quad \text{iff}\quad & \lambda_{jj}p(\vec{x}|\pi_j) \leq p(\vec{x}|\pi_N) \qquad \{j<N\}. \qquad (16)
\end{aligned}
$$

We can divide these relations by $p(\vec{x}|\pi_N)$ to obtain

$$
\begin{aligned}
\text{decide}\quad \pi_i\{i<N\}\quad \text{iff}\quad & \lambda_{ii}\mathrm{LR}_i-\lambda_{jj}\mathrm{LR}_j>0 \qquad \{i<j<N\}\\
& \text{and}\lambda_{ii}\mathrm{LR}_i>1\\
& \text{and}\lambda_{jj}\mathrm{LR}_j-\lambda_{ii}\mathrm{LR}_i\leq 0 \qquad \{j<i<N\} \qquad (17)\\
\text{decide}\quad \pi_N\quad \text{iff}\quad & \lambda_{jj}\mathrm{LR}_j\leq 1 \qquad \{j<N\}, \qquad (18)
\end{aligned}
$$

where $\mathrm{LR}_i \equiv p(\vec{x}|\pi_i)/p(\vec{x}|\pi_N)$ are the likelihood ratio decision variables used by the ideal observer.[4,5] The decision boundary hyperplanes which partition the $\vec{\mathrm{LR}} \equiv (\mathrm{LR}_1,\ldots,\mathrm{LR}_{N-1})$ decision space into the regions Z_i are thus

$$
\begin{aligned}
\lambda_{ii}\mathrm{LR}_i-\lambda_{jj}\mathrm{LR}_j &= 0 \qquad \{i<j<N\} \qquad (19)\\
\lambda_{ii}\mathrm{LR}_i &= 1 \qquad \{i<N\}. \qquad (20)
\end{aligned}
$$

Note that any of these equations, for example that defining part of the boundary between classes π_j and π_k , can be expressed as the difference of two other such equations (in this example, those defining boundaries between classes π_i and π_j, and between classes pi_i and π_k). If we require the λ_{ii} to be positive, the resulting decision rule is an ideal observer decision rule.[5] Since neither the decision variables nor the form of the decision rule depend on the particular choices of α_{ii}, we can conclude that the N-class sensitivity ROC hypersurface, obtained by allowing the λ_{ii} to take on all possible positive values, is optimal for the observer defined in Eqs. 19 and 20, in the sense that no other observer can achieve a higher sensitivity hypersurface (*i.e.*, one with a greater value of P_{NN} at a given value of $(P_{11},\ldots,P_{(N-1)(N-1)})$). The optimal observer for this performance metric is seen to be the N-class ideal observer, with its decision criteria constrained so that the boundary separating classes π_i and π_N is a hyperplane defined by $\mathrm{LR}_i = 1/\lambda_{ii}$, while the boundary separating classes π_i and π_j is a hyperplane defined by $\lambda_{ii}\mathrm{LR}_i = \lambda_{jj}\mathrm{LR}_j$.

Although an intuitive geometric understanding of this decision rule is more elusive than in the three-class case, it is at least evident that the boundaries intersect as expected; that is, the boundary separating classes π_i and π_j intersects the boundary separating classes π_i and π_k, and also intersects the boundary separating classes π_j and π_k. Note also that the number of free decision criteria has been reduced from $N^2 - N - 1$ (for the general N-class ideal observer) to $N - 1$ (as expected for a hypersurface in an N-dimensional ROC space). More importantly, comparison of Eqs. 19 and 20 with Eqs. 10–12 reveals this N-class observer to be an obvious extension from three to N classes of the observer described in the preceding section.

4. CONCLUSIONS

A fully general performance evaluation methodology for the three-class classification task has yet to be developed, a frustrating state of affairs given the great success and wide application of ROC analysis to two-class classification tasks. A primary reason for the difficulty in developing a fully general extension of ROC analysis to the three-class classification task is the rapid increase in the number of performance measurement variables and decision criteria necessary to characterize observer (in particular, ideal observer) performance. Specifically, the number of sensitivities or misclassification rates needed increases from two to six (and to $N^2 - N$ in the general case), while the number of decision criteria increases from a single decision variable threshold to a set of five mutually constrained[8] criteria (and to $N^2 - N - 1$ in the general case). In short, the complexity of the problem increases not linearly with the number of classes, but quadratically.

The motivation for the numerous proposed methods, outlined in Sec. 1, for evaluating the performance of a three-class classifier in terms of two-dimensional surfaces in three-dimensional ROC spaces (rather than the five-dimensional hypersurfaces in six-dimensional ROC spaces required by the theory) is thus quite clear. We currently lack a theoretical framework with which to judge the appropriateness of any of the proposed methods to any particular classification task. However, even if one chooses to adopt a performance evaluation metric known to provide an incomplete description of observer performance, it is still reasonable to ask what observer, if any, will achieve optimal performance with respect to that metric.

We have addressed that question in regard to measurement of an observer's performance in terms of only its sensitivities (the probabilities of correctly classifying the three, or in general N, classes of observations). Theoretically, this is clearly an incomplete measure of performance (another set of three, or in general $N^2 - 2N$, misclassification rates are necessary). Conceding this point, we consider it a nontrivial observation, derived in the preceding sections, that the observer which optimizes this limited performance metric is not one unrelated to the general ideal observer, nor an arcane special case of the ideal observer, but a special case of the ideal observer which is in a subjective sense quite simple, and which has been independently evaluated from very different perspectives by other researchers.[9,12] We find these results at once reassuring and encouraging, and hope that research into this thorny problem will continue to bear unexpected fruit.

ACKNOWLEDGMENTS

This work was supported by grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D. C. Edwards, principal investigator). Charles E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

REFERENCES

1. J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
2. C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine* **VIII**(4), pp. 283–298, 1978.
3. D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.* **31**, pp. 81–90, 2004.
4. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
5. D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.
6. C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, pp. 1–33, 1999.

7. D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks," *IEEE Trans. Med. Imag.* **24**, pp. 293–299, 2005.
8. D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.* **24**, pp. 1566–1573, 2005.
9. B. K. Scurfield, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math Psychol.* **42**, pp. 5–31, 1998.
10. D. Mossman, "Three-way ROCs," *Med. Decis. Making* **19**, pp. 78–89, 1999.
11. H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 567–578, (SPIE, Bellingham, WA), 2003.
12. X. He, C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis — I. A decision theoretic approach," *IEEE Trans. Med. Imag.*, 2005. (in review).
13. S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations: Second Edition*, Harcourt Brace Jovanovich, San Diego, CA, 1986.
14. D. C. Edwards and C. E. Metz, "Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule," in Proc. SPIE Vol. 5749 *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*, Miguel P. Eckstein and Yulei Jiang, eds., pp. 128–137, (SPIE, Bellingham, WA), 2005.
15. D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.*, 2005. (in review).

I Optimization of restricted ROC surfaces in three-class classification tasks

Optimization of restricted ROC surfaces in three-class classification tasks

Darrin C. Edwards* and Charles E. Metz

***Abstract*— We have shown previously that an N-class ideal observer achieves the optimal receiver operating characteristic (ROC) hypersurface in a Neyman-Pearson sense. Due to the inherent complexity of evaluating observer performance even in a three-class classification task, some researchers have suggested a generally incomplete but more tractable evaluation in terms of a surface plotting only the three "sensitivities." More generally, one can evaluate observer performance with a single sensitivity or misclassification probability as a function of two linear combinations of sensitivities or misclassification probabilities. We analyzed four such formulations including the "sensitivity" surface. In each case, we applied the Neyman-Pearson criterion to find the observer which achieves optimal performance with respect to each given set of "performance description variables" under consideration. In the unrestricted case, optimization with respect to the Neyman-Pearson criterion yields the ideal observer, as does maximization of the observer's expected utility. Moreover, during our consideration of the restricted cases, we found that the two optimization methods do not merely yield the same observer, but are in fact completely equivalent in a mathematical sense. Thus, for a wide variety of observers which maximize performance with respect to a restricted ROC surface in the Neyman-Pearson sense, that ROC surface can also be shown to provide a complete description of the observer's performance in an expected-utility sense.**

***Index Terms*— ROC analysis, three-class classification, ideal observer decision rules, Neyman-Pearson criterion, expected utility maximization**

I. Introduction

WE are attempting to extend the well-known observer performance evaluation methodology of receiver operating characteristic (ROC) analysis [1], [2] to classification tasks with three classes. This could conceivably be of benefit, for example, in a medical decision-making task in which a region of a patient image must be characterized as containing a malignant lesion, a benign lesion, or only normal tissue [3].

Unfortunately, a fully general extension of ROC analysis to classification tasks with more than two classes has yet to be developed. It is known that the performance of an observer in a classification task with N classes ($N \geq 2$) can be completely described by a set of $N^2 - N$ conditional error probabilities [4], [5], and that the performance of the ideal observer (that which minimizes Bayes risk [4]) is completely characterized by an ROC hypersurface in which these conditional error probabilities depend on a set of $N^2 - N - 1$ decision criteria [5]. Although analytic expressions for the ideal observer's conditional error probabilities given reasonable models for the underlying observational data have been worked out in the two-class case [6], this has not yet been accomplished in a fully general manner for tasks with three or more classes. Furthermore, we have shown that an obvious generalization of the area under the ROC curve (AUC) does not in fact yield a useful performance metric in tasks with three or more classes [7]. More recently, we showed that complicated constraining relationships exist among the decision criteria themselves for the ideal observer [8]. These constraining relationships appear to imply that it is highly unlikely that analytical expressions for the conditional error probabilities in terms of the decision criteria can be developed which are as simple to interpret as those for the two-class task [6].

Despite the difficulties just described, the potential benefits to be gained from a practical performance evaluation methodology for classification tasks with three classes have motivated a number of research groups to propose such methods. These practical methods reduce the number of degrees of freedom used to describe the observer's performance, either by implicitly leaving the remaining degrees of freedom out of the analysis, or by explicitly imposing restrictions on the form of the observer's decision rule or on the set of decision criteria used by the observer. In this work, we are concerned specifically with the latter case, and we will refer to such a model as a "restricted" performance evaluation methodology.

Scurfield evaluated an observer which used a specified decision rule with only two degrees of freedom (in general a three-class observer can have up to five degrees of freedom) by plotting a set of six (two-dimensional) surfaces in three-dimensional ROC spaces [9]. Mossman proposed plotting the surface formed only from the set of three "sensitivities" (conditional probabilities of correctly classifying observations) for an observer with two degrees of freedom, and applied this method to an observer with a specified decision rule [10]. Chan *et al.* began with an ideal observer model, and reduced the number of decision criteria from five to two by imposing explicit assumptions on the observer's decision utilities. The observer's performance was then plotted as a surface in a three-dimensional ROC space, the axes of which are the three conditional probabilities of deciding an observation to be malignant (this description of performance was also shown to be complete) [11]. He *et al.* investigated a special case of the ideal observer model which is also a special case of the decision rule proposed by Scurfield; they showed that due to the assumptions of their model, performance evaluation in terms of only the three sensitivities provides a complete

This work was supported by the US Army Medical Research and Materiel Command under Grant W81XWH-04-1-0495 (D. C. Edwards, principal investigator).

*D. C. Edwards is with the Department of Radiology, the University of Chicago, Chicago, IL, USA (e-mail: d-edwards@uchicago.edu).

C. E. Metz is with the Department of Radiology, the University of Chicago, Chicago, IL, USA.

description of this observer's performance [12]. Recently we investigated the relationships between each of these proposed decision rules and the decision rule used by the three-class ideal observer [13]; that work, however, was limited to theoretical aspects of the decision rules themselves, and did not take into account the important issue of performance measurement. The present work attempts to address this issue; it continues our analysis of the proposed decision strategies described above, specifically from the point of view of ROC analysis.

A common theme among these remarkably diverse proposed decision strategies is the idea of an "ROC surface," *i. e.*, a surface with two degrees of freedom in a three-dimensional ROC space. An appealing feature of such a construct is its visualizability: it can be plotted as readily as any elevation map, for example, in stark contrast to the fully general three-class classification task involving a hypersurface with five degrees of freedom in a six-dimensional ROC space as mentioned above.

On the other hand, it can be argued that measurement of three-class classification performance in terms of only three conditional classification rates may yield an incomplete description of observer performance; for example, a complete description of the unrestricted three-class ideal observer's performance requires six such conditional classification rates, as stated above. Acknowledging this possible incompleteness, we would like to ask whether there is any sense in which such a restricted performance evaluation method is at least well-defined. In particular, suppose we elect to measure performance in terms of an ROC surface given by a single linear combination of either sensitivities or of conditional error rates as a function of two different linear combinations of other conditional classification rates. We then ask, is there any observer decision rule, dependent on only two (rather than five) decision criteria, for which the specified ROC surface is never below (when the surface's dependent variable is a sensitivity) or never above (when the surface's dependent variable is a conditional error rate) the corresponding surface obtained for any other observer? If so, what form does this decision rule take?

In attempting to answer this question for the special cases listed above, as well as for closely related models, we applied the Neyman-Pearson criterion to find the observer which achieves optimal performance with respect to each given set of "performance description variables" (the particular set of three linear combinations of sensitivities or conditional error rates under consideration). In the unrestricted case, it is well known for $N = 2$ [4], and we showed recently for $N > 2$ [5], that optimization with respect to the Neyman-Pearson criterion yields the same observer as does maximization of the observer's expected utility (or, equivalently, minimization of Bayes's risk): namely, the ideal observer. During our consideration of the restricted cases, we found that the two optimization methods do not merely yield the same observer, but are in fact completely equivalent in a mathematical sense.

The proof of this equivalence, in the unrestricted case, is given in Sec. II. In Sec. III, we show that the equivalence holds true even in a "restricted case" such as those just mentioned — specifically, when a linear constraint is applied to the utilities used by the ideal observer to make decisions, thereby reducing the number of performance description variables required to describe the performance of the resulting observer. We then analyze four different observer decision strategies proposed recently in the literature (and known to be special cases of the three-class ideal observer [13]) in light of this result. (It should be noted that, although the restricted cases we consider are all, in fact, special cases which are in some sense "derivable" from the unrestricted model we considered previously [5], the derivations in that previous work did not consider the possibility of introducing any such constraints on the decision process.) For the reader's convenience, much of the mathematical detail of this analysis is relegated to corresponding appendices. Finally, these results are summarized, and our conclusions presented, in Secs. IV and V.

II. The Equivalence of the Neyman-Pearson and Expected Utility Optimizations

The expected utility of the decisions made by an observer in an N-class classification task can be expressed as [5]

$$\begin{aligned} E\{\mathbf{U}\} &= \sum_{i=1}^{N}\sum_{j=1}^{N} U_{i|j} P(\mathbf{d}=\pi_i, \mathbf{t}=\pi_j) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} U_{i|j} P(\mathbf{d}=\pi_i | \mathbf{t}=\pi_j) P(\mathbf{t}=\pi_j), \quad (1) \end{aligned}$$

where the labels π_1 through π_N identify the classes to which observations belong; the number $U_{i|j}$ is defined as the utility of deciding an observation belongs to class π_i given that it is actually drawn from class π_j; and the random variables $\mathbf{t}$ and $\mathbf{d}$ indicate the true class to which a randomly drawn observation belongs and the observer's decision for classifying that observation, respectively. (We use boldface type to denote statistically variable quantities.) For notational simplicity, we will write the conditional classification rate $P(\mathbf{d}=\pi_i|\mathbf{t}=\pi_j)$ as P_{ij}, and the *a priori* class membership probability $P(\mathbf{t} = \pi_i)$ as $P(\pi_i)$.

For a three-class classification task, the expected utility can be written explicitly as

$$\begin{aligned} E\{\mathbf{U}\} &= [U_{1|1}P_{11} + U_{2|1}P_{21} + U_{3|1}P_{31}]P(\pi_1) \\ &\quad + [U_{1|2}P_{12} + U_{2|2}P_{22} + U_{3|2}P_{32}]P(\pi_2) \\ &\quad + [U_{1|3}P_{13} + U_{2|3}P_{23} + U_{3|3}P_{33}]P(\pi_3). \quad (2) \end{aligned}$$

Note that the nine conditional classification rates P_{ij} appearing in this expression are not independent; for example, given the definition of conditional probability, it must be the case that $P_{11} + P_{21} + P_{31} = 1$. Thus within any pair of square brackets in (2), one of the three conditional classification rates can be eliminated, leaving an expression which depends in general on six conditional classification rates.

It can readily be shown that the observer which maximizes this expected utility is in fact the ideal observer [4], [5]. (Note that in our previous work, we demonstrated that the observer which maximizes $E_t\{U(\vec{x}, \mathbf{t})|\vec{x}\}$ is the ideal observer [5]; this is consistent with the present statement because $E\{\mathbf{U}\} = E_{\vec{x}}\{E_t[U(\vec{\mathbf{x}}, \mathbf{t})|\vec{\mathbf{x}}]\}$, and therefore maximizing the inner expectation value at each given value of $\vec{\mathbf{x}}$ will maximize $E\{\mathbf{U}\}$.) The three-class ideal observer makes decisions by partitioning a likelihood ratio decision variable plane into three regions with three intersecting lines [4], [5]. The likelihood ratios can be taken to be $\mathbf{LR}_1 \equiv p(\vec{\mathbf{x}}|\pi_1)/p(\vec{\mathbf{x}}|\pi_3)$ and $\mathbf{LR}_2 \equiv p(\vec{\mathbf{x}}|\pi_2)/p(\vec{\mathbf{x}}|\pi_3)$, ratios of the conditional probability density functions of the observational data $\vec{\mathbf{x}}$ taken as functions of that random observational data. In the notation we advocate [8], the equations for the three decision boundary lines are

$$\begin{aligned} \gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 &= \gamma_{313} - \gamma_{323} \quad (3)\\ \gamma_{131}\mathrm{LR}_1 + (\gamma_{232} - \gamma_{212})\mathrm{LR}_2 &= \gamma_{313} \quad (4)\\ (\gamma_{131} - \gamma_{121})\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 &= \gamma_{323}, \quad (5)\end{aligned}$$

which we call, respectively, the "1-*vs.*-2" line, the "1-*vs.*-3" line, and the "2-*vs.*-3" line. Here $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\pi_i)$; since the utility of a correct decision can be assumed to be greater than that of an incorrect decision, the γ_{iji} can be understood to be positive. (Note also that because (3)–(5) can be multiplied by any positive constant without changing the resulting decision boundary lines, those lines are determined by five rather than six parameters, or $N^2 - N - 1$ rather than $N^2 - N$ in general [5].) In this notation, the expression in (2) can be simplified to obtain

$$\begin{aligned} E\{\mathbf{U}\} &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3)\\ &\quad - \gamma_{121}P_{21} - \gamma_{131}P_{31}\\ &\quad - \gamma_{212}P_{12} - \gamma_{232}P_{32}\\ &\quad - \gamma_{313}P_{13} - \gamma_{323}P_{23}. \quad (6)\end{aligned}$$

An alternative method for defining "optimal performance" is in terms of the Neyman-Pearson criterion [4], [5]; the technique of satisfying the Neyman-Pearson criterion is essentially an application of an integral form of the method of Lagrange multipliers [14]. As just stated, the behavior of the ideal observer is governed by $N^2 - N - 1$ parameters, and for the present discussion we restrict our consideration of non-ideal observers to those with N^2-N-1 degrees of freedom as well. Without loss of generality, an observer's ROC hypersurface can be defined as $P_{N(N-1)}$ taken as a function of the other $N^2 - N - 1$ conditional error probabilities. (The restriction just made is then seen to be of little practical consequence: for an observer with more than $N^2 - N - 1$ degrees of freedom, one is free to consider only combinations of parameters such that $P_{N(N-1)}$ is minimized for a given set of the independent variables, reducing the number of parameters to $N^2 - N - 1$; while for an observer with fewer than $N^2 - N - 1$ degrees of freedom, it is simply the case that $P_{N(N-1)}$ is undefined for particular combinations of the independent variables.)

For the three-class classification task under consideration, the ROC hypersurface is thus given by $P_{32} = R(P_{12}, P_{13}, P_{21}, P_{23}, P_{31})$. It is reasonable to define an "optimal observer" as one that achieves the lowest possible value of P_{32} for a given set of values of $P_{12}, P_{13}, P_{21}, P_{23}, P_{31}$; this condition is known as the Neyman-Pearson criterion, and it can be shown that the observer which satisfies this criterion is in fact the ideal observer [4], [5] — *i. e.*, the same observer obtained by maximizing the expected utility in (2) or, equivalently, (6). We will not reproduce the entire derivation of that result here; it will be sufficient to outline the motivation for the Neyman-Pearson criterion.

As stated, we seek to minimize P_{32}, or, equivalently, maximize $-P_{32}$, at a particular set of values of $P_{12}, P_{13}, P_{21}, P_{23}, P_{31}$. Following the notation of Van Trees [4], we denote those particular values by α_{ij} (*e. g.*, α_{12} is the particular value of P_{12} under consideration). We then construct the function

$$F \equiv -P_{32} - \sum_{j\neq i} \lambda_{ij}(P_{ij} - \alpha_{ij}). \quad (7)$$

(The term for $i = 3$ and $j = 2$ is to be understood as excluded from the sum, here and throughout this section.) If F can be maximized over all values of the P_{ij}, and if the maximal value does not depend on the α_{ij}, then at the particular set of independent variables such that $P_{ij} = \alpha_{ij}$, the terms in the sum (the "constraints") will vanish; the maximum in F at that point will correspond simply to a minimum of P_{32} at the particular set of independent variables in question.

Since the factors λ_{ij} appearing in front of the constraints (the so-called Lagrange multipliers) are in any practical sense arbitrary, we are free to make whatever choice is convenient (effectively, this is equivalent to choosing a "scale" for F relative to P_{32} and the other conditional probabilities). Consider the change of variables

$$\gamma_{jij} \equiv \gamma_{232}\lambda_{ij}, \quad (8)$$

where γ_{232} is in turn defined as some arbitrary positive constant (*cf.* the statement after (5) that the set of six values of γ_{jij} could be reduced to five by multiplying by any convenient positive constant). The values of γ_{232} and the other γ_{jij} are here assumed to be positive; although the λ_{ij} are, as just stated, effectively arbitrary, we will be able to show shortly that this restriction to positive values does not result in a loss of generality.

With this substitution, the Neyman-Pearson function can be rewritten as

$$\begin{aligned} \gamma_{232}F &= -\gamma_{232}P_{32} - \sum_{j\neq i} \gamma_{jij}(P_{ij} - \alpha_{ij})\\ &= E\{\mathbf{U}\} - \sum U_{i|i}P(\pi_i) + \sum_{j\neq i} \gamma_{jij}\alpha_{ij}, \quad (9)\end{aligned}$$

i. e., the expression for expected utility plus constant terms independent of the observer's decision rule (which determines the P_{ij}). In this form, the fact that maximization of expected utility and satisfaction of the Neyman-Pearson criterion both yield the ideal observer is seen to be not merely an elegant convenience, but a necessary consequence of the mathematical equivalence of the two methods. It is also worth noting that, by replacing γ_{232} with the more general $\gamma_{(N-1)N(N-1)}$ and removing the implicit restrictions on i and j to $(1, 2, 3)$, the equivalence in (9) is seen to hold for classification tasks with an arbitrary number of classes, not just three.

We can now also justify the claim just made that the Lagrange multipliers λ_{ij} can be restricted to positive values without loss of generality (assuming γ_{232} to be an arbitrary positive constant). In the context of expected utility, a negative value of λ_{ij} or, equivalently, of γ_{jij} would correspond to an incorrect decision having a utility greater than that of the corresponding correct decision. This possibility (equivalent in a two-class classification task to an ROC operating point "below the guessing line") can, at least in the context of the ideal observer, be ignored as being "perverse." (A zero value of γ_{jij} corresponds to an incorrect decision having a utility exactly equal to that of the corresponding correct decision. Although we choose to ignore this situation in the general case, a model in which some of the γ_{jij} are set to zero without contradiction is considered in Sec. III-B.)

III. Restricted ROC Surfaces

A. Theoretical Considerations

In the Introduction, it was pointed out that the complexity of the three-class classification task has so far hindered the development of a fully general extension of ROC analysis to this task. As a result, many researchers have proposed simplified or restricted performance evaluation strategies; a number of these, also mentioned in the Introduction, involve the imposition of linear "constraints" on the utilities used by the ideal observer to make decisions. (In previous work, we examined the relationship between those proposed decision rules and the decision rule used by the ideal observer, without explicit regard to performance evaluation issues [13].) The practical effect of these constraints, as will be shown in more detail in the remainder of this section, is to reduce the number of performance description variables (the sensitivities or conditional error rates) needed to describe the observer's performance. In this section, we will first demonstrate that the equivalence between expected utility maximization and optimization through the Neyman-Pearson criterion also holds when arbitrary linear constraints are placed on the decision utilities; this result is shown to hold true for classification tasks with an arbitrary number of classes. We will then illustrate this equivalence explicitly by considering four proposed restricted three-class models.

Consider a simple linear constraint on the decision utilities of the form $U_{i|j} = U_{k|l}$. In the special case $i = j = l$, we clearly have $\gamma_{iki} = (U_{i|i} - U_{k|i})P(\pi_i) = 0$. Similarly, any linear constraint on the utilities $U_{i|j}$ can be reexpressed as a linear constraint on the γ_{iji} parameters, which we can write as

$$\gamma_{iji} = \sum_{\substack{k \neq l \\ (k,l) \neq (i,j)}} v_{lk} \gamma_{klk}, \tag{10}$$

where the v_{lk} are a set of constants determining the constraint.

Substituting (10) into (6) allows us to write

$$\begin{aligned} E_{\vec{v}}\{\mathbf{U}\} &= \sum U_{i|i} P(\pi_i) \\ &\quad - \{\ldots + \gamma_{klk}(P_{lk} + v_{lk}P_{ji}) + \ldots\}. \end{aligned} \tag{11}$$

(Here the subscript $\vec{v}$ on the expectation operator denotes the restriction imposed on the utilities *via* the v_{lk}, and not a random variable over which the expectation is taken.) Note that γ_{iji} no longer appears in the expression for expected utility, which now depends on a set of $N^2 - N - 1$ generalized performance description variables (GPDVs) — *i.e.*, the expressions $P_{lk} + v_{lk}P_{ji}$. In general, of course, these may not have any obvious practical interpretation in terms of the performance of the observer (hence the use of the word "generalized"). For non-negative values of v_{lk}, however, it is at least the case that a weighted sum of sensitivities will still behave in some sense like a sensitivity (higher values for a given observer are better than lower ones), and a weighted sum of conditional error rates still behaves like a conditional error rate (lower values are better than higher ones). This should be regarded as a practical rather than theoretical consideration, and it is in some sense an obligation of a proposed restricted method that the actual GPDVs involved be justifiable (or at least interpretable). This will be attempted in the remainder of this section during consideration of the four special cases referred to above.

For the moment, note that if we construct a Neyman-Pearson function $F_{\vec{v}}$ from the remaining $N^2 - N - 1$ GPDVs, analogous to (7), the result, after a suitable selection of the λ_{lk} parameters, will again be a complete equivalence between $F_{\vec{v}}$ and $E_{\vec{v}}\{\mathbf{U}\}$. That is, the expressions will be equal to within a positive scale factor and an additive term independent of the observer's decision rule. It remains only to note that an arbitrary number of such linear constraints can be further imposed (up to a total of $N^2 - N - 1$, in order to be left with at least one GPDV) with equivalence continuing to hold. In the next four subsections, we consider three-class classification tasks in which three constraints are imposed on the utilities, leaving a set of three GPDVs (*i. e.*, an ROC surface with two degrees of freedom in a three-dimensional ROC space).

Before turning to those special cases, however, it is perhaps worth summarizing the results of the preceding paragraphs. Briefly, if one imposes particular constraints on the behavior of an N-class ideal observer, the resulting expected utility for that constrained observer will depend on fewer than $N^2 - N$ GPDVs. Description of the constrained observer's performance in terms of this reduced number of GPDVs is therefore complete from the point of view of expected utility. Furthermore, given the mathematical equivalence of $F_{\vec{v}}$ and $E_{\vec{v}}\{\mathbf{U}\}$ just demonstrated, the performance of the observer which maximizes $F_{\vec{v}}$ is also completely described by the same reduced set of GPDVs.

B. The Chan et al. Observer

Chan *et al.* consider a three-class classification task in which class π_1 represents "benign," class π_2 "normal," and class π_3 "malignant" observations (*e. g.*, for structures evident in a medical image) [11]. They simplify the expression in (2) by restricting all values of utility to lie between $U_{\min}$ and $U_{\max}$; by setting the "correct decision" utilities $U_{1|1}$, $U_{2|2}$, and $U_{3|3}$ to be $U_{\max}$; the "missed malignancy" utilities $U_{1|3}$ and $U_{2|3}$ to be $U_{\min}$; and the utilities for incorrect decisions not involving malignancies $U_{1|2}$ and $U_{2|1}$ to be $U_{\max}$. The remaining "false-positive" utilities $U_{3|1}$ and $U_{3|2}$ are free to vary in the range $[U_{\min}, U_{\max}]$. In our notation, this corresponds to imposing the three constraints $\gamma_{121} = 0$, $\gamma_{212} = 0$, and $\gamma_{313} = \gamma_{323}$. (The remaining condition $\gamma_{313} = (U_{\max} - U_{\min})P(\pi_3)$ is not an additional constraint — in the sense of restricting the form of the observer's decision rule — but merely determines the scale

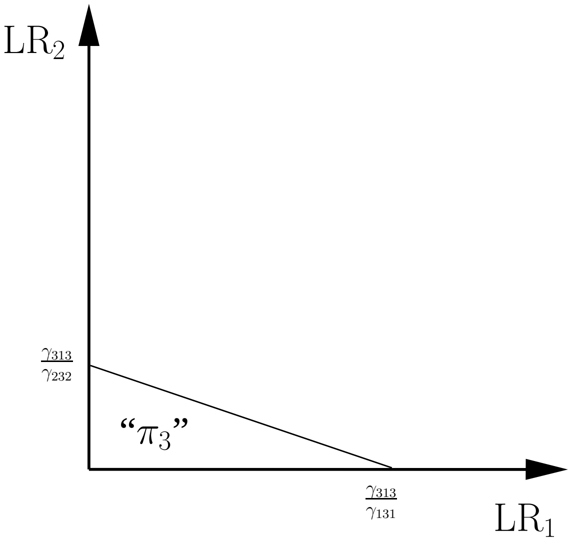

LR_2
$\frac{\gamma_{313}}{\gamma_{232}}$
"π_3"
$\frac{\gamma_{313}}{\gamma_{131}}$
LR_1

Fig. 1. The decision strategy investigated by Chan *et al.*, which is a special case of the ideal observer decision strategy. Observations in the unlabeled region are decided "not π_3," *i. e.*, either "π_1" or "π_2".

of the remaining parameters as explained in the text following (5).)

With these assumptions, the expression for expected utility is reduced to

$$\begin{aligned} E\{\mathbf{U}_{\text{Chan}}\} &= U_{\max} \\ &\quad - \gamma_{131}P_{31} - \gamma_{232}P_{32} \\ &\quad - \gamma_{313}(P_{13} + P_{23}) \\ &= U_{\max} - \gamma_{313} \\ &\quad - \gamma_{131}P_{31} - \gamma_{232}P_{32} + \gamma_{313}P_{33}, \quad (12) \end{aligned}$$

since $P_{13} + P_{23} = 1 - P_{33}$ (again, note that γ_{131}, γ_{232}, and γ_{313} are dependent on only two free parameters $U_{3|1}$ and $U_{3|2}$). As Chan *et al.* point out [11], this expression depends on three rather than six GPDVs, namely P_{31}, P_{32}, and P_{33}. These three rates are used to construct the ROC space in which they analyze the performance of their observer. That observer in turn is the special case of the ideal observer obtained by imposing the above constraints on the decision utilities $U_{i|j}$ or, equivalently, on the parameters γ_{jij}.

Although we have found it useful to assume the quantities γ_{jij} to be strictly positive, this is not a fundamental requirement, and Chan *et al.* indeed allow some of them (*e. g.*, γ_{121}) to be zero (consistent with the constraints they place on the $U_{i|j}$ as described above). They obtain the resulting ideal observer decision lines

$$\begin{aligned} 0\mathrm{LR}_1 - 0\mathrm{LR}_2 &= 0 & \{\text{“1-}\textit{vs.}\text{-2”}\} \quad (13) \\ \gamma_{131}\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 &= \gamma_{313} & \{\text{“1-}\textit{vs.}\text{-3”}\} \quad (14) \\ \gamma_{131}\mathrm{LR}_1 + \gamma_{232}\mathrm{LR}_2 &= \gamma_{313} & \{\text{“2-}\textit{vs.}\text{-3”}\}, \quad (15) \end{aligned}$$

which actually correspond to a single line (as the first is undefined and the remaining two are degenerate). This decision strategy is illustrated in Fig. 1.

In summary, Chan *et al.* begin with a three-class ideal observer model, impose particular constraints on the decision utilities in that model, and then determine, based on those constraints, both the resulting form of the special case of the ideal observer and the conditional classification rates appropriate to measuring its performance. We now wish to pose a question from a different point of view: suppose one chooses to measure arbitrary (*i. e.*, not necessarily ideal) observer performance only in terms of the conditional classification rates P_{33}, P_{31}, and P_{32}, ignoring the other rates. For any observer, we can construct an ROC surface with P_{33} as a function of P_{31} and P_{32}. (For an observer with more than two degrees of freedom in its decision strategy, one can simply define the surface to be the maximum value of P_{33} achievable at any given (P_{31}, P_{32}) pair.) What observer, if any, will achieve optimal performance with respect to this surface?

We seek to maximize P_{33} at a particular point $(P_{31} = \alpha_{31}, P_{32} = \alpha_{32})$ in the domain of the given ROC space. Another way of stating this is to consider P_{33}, P_{31}, and P_{32} as functionals of the observer's decision rule; we seek to maximize P_{33} subject to the constraints $P_{31} = \alpha_{31}$ and $P_{32} = \alpha_{32}$. To find this maximum, we define a function

$$F_{\text{Chan}} \equiv P_{33} - \lambda_{31}(P_{31} - \alpha_{31}) - \lambda_{32}(P_{32} - \alpha_{32}), \qquad (16)$$

where λ_{31} and λ_{32} are free parameters (the so-called Lagrange multipliers). Note that maximizing F_{Chan} at the particular point $(P_{31} = \alpha_{31}, P_{32} = \alpha_{32})$ is equivalent to maximizing P_{33} at that point; if the maxima for arbitrary points (P_{31}, P_{32}) are achieved by a single decision rule independent of α_{31} and α_{32}, the resulting surface will be the desired optimal surface.

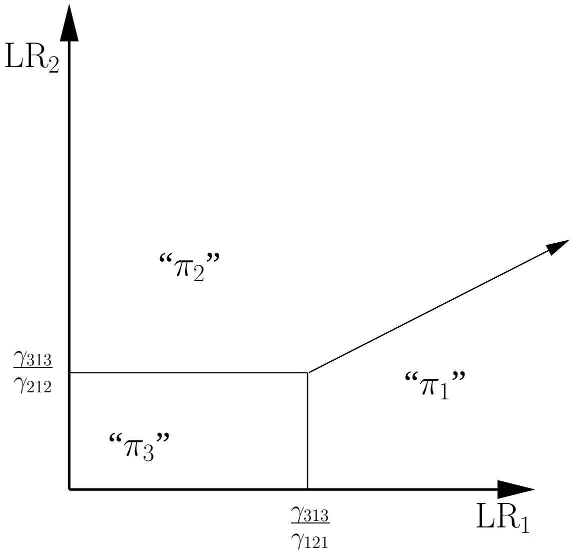
The functional in (16) is maximized in App. A. The boundary lines which partition the $(\mathbf{LR}_1, \mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are found to be

$$\begin{aligned} 0\mathrm{LR}_1 - 0\mathrm{LR}_2 &= 0 & \{\text{“1-}\textit{vs.}\text{-2”}\} \quad (17) \\ \lambda_{31}\mathrm{LR}_1 + \lambda_{32}\mathrm{LR}_2 &= 1 & \{\text{“1-}\textit{vs.}\text{-3”}\} \quad (18) \\ \lambda_{31}\mathrm{LR}_1 + \lambda_{32}\mathrm{LR}_2 &= 1 & \{\text{“2-}\textit{vs.}\text{-3”}\}. \quad (19) \end{aligned}$$

If we require λ_{31} and λ_{32} to be positive, and then define the quantities $\gamma_{131} \equiv \gamma_{313}\lambda_{31}$ and $\gamma_{232} \equiv \gamma_{313}\lambda_{32}$ for a positive constant γ_{313}, the resulting decision strategy is found to be identical to that stated in (13)–(15). The special case of the ideal observer proposed by Chan *et al.*, whose performance depends only on the conditional classification rates P_{33}, P_{31}, and P_{32} by (12), is indeed the observer which obtains optimal performance with respect to this set of conditional classification rates. By the argument at the end of Sec. III-A, this description of the constrained observer's performance is complete.

C. The He et al. Observer

He *et al.* also begin with a three-class ideal observer model and thus with the expression for expected utility given in (2); the classification task of interest to them is to distinguish normal, infarcted, and ischemic tissue based on myocardial perfusion SPECT [12]. They simplify this expression by requiring that the two possible incorrect classifications of observations actually from a given class be equal. That is, $U_{2|1} = U_{3|1}$, $U_{1|2} = U_{3|2}$, and $U_{1|3} = U_{2|3}$. These can immediately be expressed as the (linear) constraints $\gamma_{121} =$



LR_2
"π_2"
"π_1"
"π_3"
$\frac{\gamma_{313}}{\gamma_{212}}$
$\frac{\gamma_{313}}{\gamma_{121}}$
LR_1

Fig. 2. The decision strategy investigated by He *et al.*, which is a special case of the ideal observer decision strategy, and which can also be shown to be a special case of the Scurfield observer in which the decision variables used are the logarithms of the likelihood ratios $(\mathbf{LR}_1, \mathbf{LR}_2)$ of the observational data.

γ_{131}, $\gamma_{212} = \gamma_{232}$, and $\gamma_{313} = \gamma_{323}$. The expression for expected utility is thereby reduced, in our notation, to

$$\begin{aligned} E\{\mathbf{U}_{\text{He}}\} &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3) \\ &\quad - \gamma_{121}(P_{21} + P_{31}) \\ &\quad - \gamma_{212}(P_{12} + P_{32}) \\ &\quad - \gamma_{313}(P_{12} + P_{32}) \\ &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3) \\ &\quad - (\gamma_{121} + \gamma_{212} + \gamma_{313}) \\ &\quad + \gamma_{121}P_{11} + \gamma_{212}P_{22} + \gamma_{313}P_{33}. \end{aligned} \tag{20}$$

As He *et al.* point out [12], this expression depends on only the three "sensitivities" P_{11}, P_{22}, and P_{33}, rather than six GPDVs. The three sensitivities are used to construct the ROC space (equivalent to that proposed by Mossman [10]) in which they analyze the performance of their observer. That observer in turn is the special case of the ideal observer obtained by imposing the above constraints on the decision utilities $U_{i|j}$ or, equivalently, on the parameters γ_{jij} .

Applying the stated constraints on the utilities to the ideal observer decision boundary lines given in (3)–(5) yields

$$\begin{aligned} \gamma_{121}\mathrm{LR}_1 - \gamma_{212}\mathrm{LR}_2 &= 0 && (21) \\ \gamma_{121}\mathrm{LR}_1 &= \gamma_{313} && (22) \\ \gamma_{212}\mathrm{LR}_2 &= \gamma_{313}. && (23) \end{aligned}$$

This decision strategy is illustrated in Fig. 2. We have recently shown [13] that this decision strategy is a special case of that proposed by Scurfield [9] when the decision variables used by the Scurfield observer are the logarithms of the likelihood ratios of the observational data.

We now consider evaluating the performance of an arbitrary observer in the ROC space constructed only from the observer's sensitivities (*i. e.*, P_{11}, P_{22}, and P_{33}). Without loss of generality, we can define such an observer's ROC surface as P_{33} considered as a function of P_{11} and P_{22}; to find the optimal observer with respect to this restricted performance evaluation method, we apply the Neyman-Pearson criterion to maximize P_{33} subject to the constraints $(P_{11} = \alpha_{11}, P_{22} = \alpha_{22})$. We define the function

$$F_{\text{He}} \equiv P_{33} + \lambda_{11}(P_{11} - \alpha_{11}) + \lambda_{22}(P_{22} - \alpha_{22}), \tag{24}$$

where λ_{11} and λ_{22} are again the Lagrange multipliers.

The functional in (24) is maximized in App. B. The boundary lines which partition the $(\mathbf{LR}_1, \mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are found to be

$$\begin{aligned} \lambda_{11}\mathrm{LR}_1 - \lambda_{22}\mathrm{LR}_2 &= 0 && \{\text{"1-}vs.\text{-2"}\} && (25) \\ \lambda_{11}\mathrm{LR}_1 &= 1 && \{\text{"1-}vs.\text{-3"}\} && (26) \\ \lambda_{22}\mathrm{LR}_2 &= 1 && \{\text{"2-}vs.\text{-3"}\}. && (27) \end{aligned}$$

If we require λ_{11} and λ_{22} to be positive, and define the quantities $\gamma_{121} \equiv \gamma_{313}\lambda_{11}$ and $\gamma_{212} \equiv \gamma_{313}\lambda_{22}$ for some arbitrary positive constant γ_{313}, then the resulting decision strategy is found to be identical to that stated in (21)–(23). The special case of the ideal observer proposed by He *et al.*, whose performance depends only on the conditional classification rates P_{11}, P_{22}, and P_{33} by (20), is indeed the observer which obtains optimal performance with respect to this set of conditional classification rates. By the argument at the end of Sec. III-A, this description of the constrained observer's performance is complete.

D. The Scurfield Observer (Likelihood Ratio)

In the preceding two sections, we considered decision strategies that have been proposed by other researchers as special cases of the three-class ideal observer decision strategy. That is, particular constraints were explicitly imposed in the work cited on the decision utilities used by the ideal observer. The remaining two decision strategies we consider in the present work are special cases of a decision strategy proposed by Scurfield [9] which was not claimed to be generally related to the ideal observer; specifically, Scurfield specified the decision boundary lines used by the observer, but made no assumptions concerning the observer's two decision variables.

We showed recently [13] that if particular forms of the observer's decision variables related to the likelihood ratios of the observational data are chosen, then the resulting decision strategies can be shown to be special cases of the ideal observer decision strategy. One such special case is the observer analyzed by He *et al.* [12], discussed in Sec. III-C, in which the decision variables used by the Scurfield observer are the logarithms of the likelihood ratios. Two other such special cases are the Scurfield observer with the likelihood ratios themselves as decision variables, which we consider in this section; and that with the *a posteriori* class membership probabilities used as decision variables, considered in Sec. III-E. A minor difference from the preceding two sections is that we must determine the implicit constraints on the ideal observer's utilities from the known form of the decision rule, rather than the other way around.

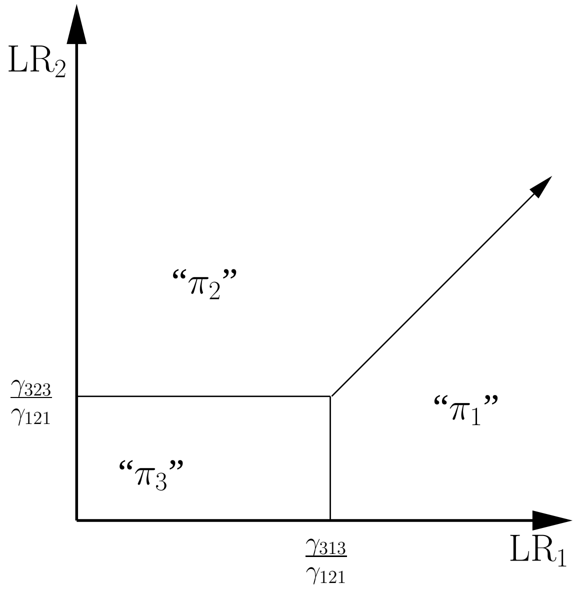

LR_2
"π_2"
$\frac{\gamma_{323}}{\gamma_{121}}$
"π_1"
"π_3"
$\frac{\gamma_{313}}{\gamma_{121}}$
LR_1

Fig. 3. A special case of the decision strategy investigated by Scurfield, in which the decision variables used are the likelihood ratios $(\mathbf{LR}_1, \mathbf{LR}_2)$ of the observational data.

The general Scurfield observer makes decisions by partitioning a decision variable plane $(\mathbf{y}_1, \mathbf{y}_2)$ into three regions *via* the decision boundary lines

$$\begin{aligned} y_1 - y_2 &= \gamma_1 - \gamma_2 && (28)\\ y_1 &= \gamma_1 && (29)\\ y_2 &= \gamma_2, && (30) \end{aligned}$$

where γ_1 and γ_2 are parameters upon which the observer's performance depends (roughly equivalent to the decision criterion of a two-class classifier) [9]. When the decision variables are themselves the likelihood ratios $(\mathbf{LR}_1, \mathbf{LR}_2)$, this becomes in our notation

$$\begin{aligned} \gamma_{121}\mathrm{LR}_1 - \gamma_{121}\mathrm{LR}_2 &= \gamma_{313} - \gamma_{323} && (31)\\ \gamma_{121}\mathrm{LR}_1 &= \gamma_{313} && (32)\\ \gamma_{121}\mathrm{LR}_2 &= \gamma_{323}. && (33) \end{aligned}$$

(Compare (28)–(30) with (3)–(5), and note that in order for the "1-*vs.*-2" line to have unit slope, it must be the case that $\gamma_{121} = \gamma_{212}$. Alternatively, after making the assignments $y_1 \equiv \mathrm{LR}_1$, $y_2 \equiv \mathrm{LR}_2$ in (28)–(30), one is free to multiply all three equations by a postive constant γ_{121}.) This decision strategy is illustrated in Fig. 3.

The relations $\gamma_{121} = \gamma_{131}$ and $\gamma_{212} = \gamma_{232}$ evident from the above equations immediately give the constraints on the decision utilities $U_{2|1} = U_{3|1}$ and $U_{1|2} = U_{3|2}$. Furthermore, the constraint $\gamma_{121} = \gamma_{212}$ implies $(U_{1|1} - U_{2|1})P(\pi_1) = (U_{2|2} - U_{1|2})P(\pi_2)$. (Recall from Sec. II that $\gamma_{iji} \equiv (U_{i|i} - U_{j|i})P(\pi_i)$.) This allows us to simplify the expression for expected utility in (2) to yield

$$\begin{aligned} E\{\mathbf{U}_{\text{Scfd:LR}}\} &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3)\\ &\quad - \gamma_{121}(P_{21} + P_{31}) - \gamma_{121}(P_{12} + P_{32})\\ &\quad - \gamma_{313}P_{13} - \gamma_{323}P_{23}\\ &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3)\\ &\quad - 2\gamma_{121} + \gamma_{121}(P_{11} + P_{22})\\ &\quad - \gamma_{313}P_{13} - \gamma_{323}P_{23}. && (34) \end{aligned}$$

This expression for the observer's expected utility depends on only three GPDVs: P_{13} and P_{23}, which are just the misclassification rates for observations actually drawn from class π_3; and $P_{11} + P_{22}$, which may be regarded as the "total sensitivity" for observations actually drawn from classes π_1 and π_2 (ignoring the *a priori* rates for such observations).

We now consider evaluating the performance of an arbitrary observer in an ROC-like space constructed from the quantities $P_{11} + P_{22}$, P_{13}, and P_{23}. We will define the ROC-like surface used to evaluate observer performance as the first quantity considered as a function of the two misclassification rates. To find the optimal observer with respect to this restricted performance evaluation method, we apply the Neyman-Pearson criterion to maximize $P_{11} + P_{22}$ subject to the constraints $(P_{13} = \alpha_{13}, P_{23} = \alpha_{23})$. We define the function

$$\begin{aligned} F_{\text{Scfd:LR}} &\equiv P_{11} + P_{22} - \lambda_{13}(P_{13} - \alpha_{13})\\ &\quad - \lambda_{23}(P_{23} - \alpha_{23}), && (35) \end{aligned}$$

where λ_{13} and λ_{23} are the Lagrange multipliers.

The functional in (35) is maximized in App. C. The boundary lines which partition the $(\mathbf{LR}_1, \mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are found to be

$$\begin{aligned} \mathrm{LR}_1 - \mathrm{LR}_2 &= \lambda_{13} - \lambda_{23} && \{\text{"1-}\textit{vs.}\text{-2"}\} && (36)\\ \mathrm{LR}_1 &= \lambda_{13} && \{\text{"1-}\textit{vs.}\text{-3"}\} && (37)\\ \mathrm{LR}_2 &= \lambda_{23} && \{\text{"2-}\textit{vs.}\text{-3"}\}. && (38) \end{aligned}$$

If we require λ_{13} and λ_{23} to be positive, and define the quantities $\gamma_{313} \equiv \gamma_{121}\lambda_{13}$ and $\gamma_{323} \equiv \gamma_{121}\lambda_{23}$ for some arbitrary positive constant γ_{121}, then the resulting decision strategy is found to be identical to that stated in (31)–(33). This special case of the observer proposed by Scurfield, which we have shown to be a special case of the ideal observer [13], has a performance that depends only on the GPDVs $P_{11}+P_{22}$, P_{13}, and P_{23} by (34). This is indeed the observer which obtains optimal performance with respect to this set of quantities related to the conditional classification rates. By the argument at the end of Sec. III-A, this description of the constrained observer's performance is complete.

E. The Scurfield Observer (a posteriori Class Probability)

Equations (28)–(30) in Sec. III-D give the equations for the decision boundary lines of the general Scurfield observer. If we now use two of the *a posteriori* class membership probabilities, such as $P(\pi_1|\mathbf{\vec{x}})$ and $P(\pi_2|\mathbf{\vec{x}})$, as the decision variables, the equations become

$$\begin{aligned} P(\pi_1|\vec{x}) - P(\pi_2|\vec{x}) &= \gamma_1 - \gamma_2 && (39)\\ P(\pi_1|\vec{x}) &= \gamma_1 && (40)\\ P(\pi_2|\vec{x}) &= \gamma_2, && (41) \end{aligned}$$

with $0 \leq \gamma_1 \leq 1$ and $0 \leq \gamma_2 \leq 1$. (Note that $P(\pi_3|\vec{x}) = 1 - P(\pi_1|\vec{x}) - P(\pi_2|\vec{x})$, meaning this third probability is not needed as an independent decision variable; the particular

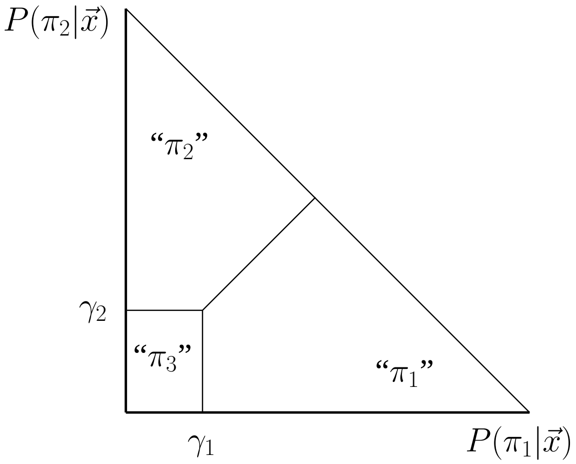


Fig. 4. A special case of the decision strategy investigated by Scurfield, in which the decision variables used are the *a posteriori* class membership probabilities $P(\pi_1|\mathbf{\vec{x}})$ and $P(\pi_2|\mathbf{\vec{x}})$ of the observational data.

choice of which two probabilities to use is of course arbitrary.) This decision strategy, which we have shown recently to be a special case of the ideal observer decision strategy [13], is illustrated in Fig. 4.

We can reexpress the above equations in terms of likelihood ratios by exploiting the relation

$$\begin{aligned} P(\pi_i|\vec{x}) &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{p(\vec{x})} \\ &= \frac{k_i \mathrm{LR}_i}{1 + k_1\mathrm{LR}_1 + k_2\mathrm{LR}_2}, \end{aligned} \tag{42}$$

where the second equation is obtained by dividing the numerator and denominator of the first by $p(\vec{x}|\pi_3)P(\pi_3)$, and where $k_i \equiv P(\pi_i)/P(\pi_3)$. The equations for the decision boundary lines become

$$\begin{aligned} \frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 - \frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 &= (\gamma_1 - \gamma_2)\left(1 + \frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1\right. \\ &\qquad \left. + \frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2\right) \qquad (43) \\ \frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1 &= \gamma_1\left(1 + \frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1\right. \\ &\qquad \left. + \frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2\right) \qquad (44) \\ \frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 &= \gamma_2\left(1 + \frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1\right. \\ &\qquad \left. + \frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2\right), \qquad (45) \end{aligned}$$

which can in turn be simplified to yield

$$\begin{aligned} &[1 - (\gamma_1 - \gamma_2)]P(\pi_1)\mathrm{LR}_1 \\ &\quad -[1 + (\gamma_1 - \gamma_2)]P(\pi_2)\mathrm{LR}_2 \\ &\qquad = (\gamma_1 - \gamma_2)P(\pi_3) \qquad (46) \\ (1-\gamma_1)P(\pi_1)\mathrm{LR}_1 - \gamma_1 P(\pi_2)\mathrm{LR}_2 &= \gamma_1 P(\pi_3) \qquad (47) \\ -\gamma_2 P(\pi_1)\mathrm{LR}_1 + (1-\gamma_2)P(\pi_2)\mathrm{LR}_2 &= \gamma_2 P(\pi_3). \qquad (48) \end{aligned}$$

Although the above equations for the decision boundary lines are much more complicated than those of the previous three sections, we can still relate the parameters γ_1 and γ_2 to the decision rule parameters of (3)–(5) to obtain constraints on them and, consequently, on the utilities $U_{i|j}$. For example, comparison of (47) with (4) gives

$$\gamma_{232} - \gamma_{212} = \frac{-P(\pi_2)}{P(\pi_3)}\gamma_{313}, \tag{49}$$

which can also be expressed in terms of the utilities as $-(U_{1|2} - U_{3|2}) = U_{3|3} - U_{1|3}$. Similarly, comparison of (48) and (5) gives

$$\gamma_{131} - \gamma_{121} = \frac{-P(\pi_1)}{P(\pi_3)}\gamma_{323}, \tag{50}$$

which can be expressed in terms of the utilities as $-(U_{2|1} - U_{3|1}) = U_{3|3} - U_{2|3}$. Finally, we add the first two coefficients of (46) and then compare with (3) to obtain

$$\frac{\gamma_{121}}{P(\pi_1)} - \frac{\gamma_{212}}{P(\pi_2)} = \frac{-2(\gamma_{313} - \gamma_{323})}{P(\pi_3)}, \tag{51}$$

which can be expressed in terms of the utilities as $(U_{1|1} - U_{2|1}) - (U_{2|2} - U_{1|2}) = -2(U_{2|3} - U_{1|3})$. Note that the remaining terms in (46)–(48) involving γ_1 or γ_2 are simply differences of terms already considered, and would thus yield no further constraints on the utilities.

We can now impose constraints (49), (50), and (51) on the general expression (2) for expected utility to obtain the expected utility for this observer:

$$\begin{aligned} E\{\mathbf{U}_{\text{Scfd:AP}}\} &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3) \\ &\quad - \gamma_{121}P_{21} - \left[\frac{P(\pi_2)}{P(\pi_1)}\gamma_{121}\right. \\ &\quad \left. + \frac{2P(\pi_2)}{P(\pi_3)}(\gamma_{313} - \gamma_{323})\right] P_{12} \\ &\quad - \left[\gamma_{121} - \frac{P(\pi_1)}{P(\pi_3)}\gamma_{323}\right] P_{31} \\ &\quad - \left[\frac{P(\pi_2)}{P(\pi_1)}\gamma_{121} + \frac{P(\pi_2)}{P(\pi_3)}\gamma_{313}\right. \\ &\quad \left. - \frac{2P(\pi_2)}{P(\pi_3)}\gamma_{323}\right] P_{32} \\ &\quad - \gamma_{313}P_{13} - \gamma_{323}P_{23} \\ &= U_{1|1}P(\pi_1) + U_{2|2}P(\pi_2) + U_{3|3}P(\pi_3) \\ &\quad - \frac{\gamma_{121}}{P(\pi_1)}\left[P(\pi_1)(P_{21} + P_{31})\right. \\ &\quad \left. + P(\pi_2)(P_{12} + P_{32})\right] \\ &\quad - \frac{\gamma_{313}}{P(\pi_3)}\left[2P(\pi_2)P_{12} + P(\pi_2)P_{32}\right. \\ &\quad \left. + P(\pi_3)P_{13}\right] \\ &\quad - \frac{\gamma_{323}}{P(\pi_3)}\left[-2P(\pi_2)(P_{12} + P_{32})\right. \\ &\quad \left. - P(\pi_1)P_{31} + P(\pi_3)P_{23}\right]. \end{aligned} \tag{52}$$

This can in turn be simplified slightly using the definition of conditional probability to yield

$$\begin{aligned}
E\{\mathbf{U}_{\text{Scfd:AP}}\} &= U_{1|1}P(\pi_1)+U_{2|2}P(\pi_2)+U_{3|3}P(\pi_3) \\
&\quad -\frac{P(\pi_1)+P(\pi_2)}{P(\pi_1)}\gamma_{121}+\frac{2P(\pi_2)}{P(\pi_3)}\gamma_{323} \\
&\quad +\frac{\gamma_{121}}{P(\pi_1)}\left[P(\pi_1)P_{11}+P(\pi_2)P_{22}\right] \\
&\quad -\frac{\gamma_{313}}{P(\pi_3)}\left[2P(\pi_2)P_{12}+P(\pi_2)P_{32}\right. \\
&\quad \left.+P(\pi_3)P_{13}\right] \\
&\quad -\frac{\gamma_{323}}{P(\pi_3)}\left[2P(\pi_2)P_{22}-P(\pi_1)P_{31}\right. \\
&\quad \left.+P(\pi_3)P_{23}\right]. \qquad (53)
\end{aligned}$$

As was the case for the decision strategies of the preceding three sections, the expected utility of this observer (and thus its performance, as it too is a special case of the ideal observer) depends on only three GPDVs, namely the quantities in square brackets in (53).

The first GPDV, being a weighted sum of "sensitivities" with positive weights, is immediately seen to be quite suitable for the dependent variable of an ROC surface — a higher value of this quantity is clearly preferable to a lower one. (Indeed, $P(\pi_1)P_{11}+P(\pi_2)P_{22}$ has an intuitive interpretation as the probability of a randomly drawn observation being both (i) from either class π_1 or π_2 and also (ii) correctly classified as such. Compare the corresponding quantity $P_{11}+P_{22}$ from Sec. III-D, which is technically not even a probability.) The other two GPDVs in (53) discourage any such straightforward interpretation, but this is perhaps to be expected: the pleasantly symmetric form of the Scurfield decision rule of (28)–(30) in this case holds in the $(P(\pi_1|\vec{\mathbf{x}}), P(\pi_2|\vec{\mathbf{x}}))$ decision variable plane; due to the complexity of the transformation in (42), this symmetry will be lost in the likelihood ratio decision variable plane, and the expression for expected utility will be correspondingly opaque. (Despite this complexity, it is worth emphasizing that the Scurfield decision rule, for arbitrary choice of the decision variables, has the advantage that it can be proven rigorously that the volume under any of the conventional ROC surfaces proposed by Scurfield is equal to the probability of a particular outcome of a three-alternative forced choice experiment [9]. Although it is possible that a different choice of decision *rule* would yield a more "intuitive" triple of GPDVs than that given in (53), we have considered it worthwhile to investigate the consequences of the Scurfield decsion rule for three very natural choices of decision *variable* — namely, the log-likelihood ratios investigated by He; the likelihood ratios themselves; and the *a posteriori* class membership probabilities.)

In any case, we now consider evaluating the performance of an arbitrary observer in an ROC-like space constructed from the quantities $P(\pi_1)P_{11}+P(\pi_2)P_{22}$, $2P(\pi_2)P_{12}+P(\pi_2)P_{32}+P(\pi_3)P_{13}$, and $2P(\pi_2)P_{22}-P(\pi_1)P_{31}+P(\pi_3)P_{23}$. We will define the ROC-like surface used to evaluate observer performance as the first quantity considered as a function of the other two. To find the optimal observer with respect to this restricted performance evaluation method, we apply the Neyman-Pearson criterion to maximize $P(\pi_1)P_{11}+P(\pi_2)P_{22}$ subject to the constraints $2P(\pi_2)P_{12}+P(\pi_2)P_{32}+P(\pi_3)P_{13}=\alpha_1$ and $2P(\pi_2)P_{22}-P(\pi_1)P_{31}+P(\pi_3)P_{23}=\alpha_2$. We define the function

$$\begin{aligned}
F_{\text{Scfd:AP}} &\equiv P(\pi_1)P_{11}+P(\pi_2)P_{22} \\
&\quad -\lambda_1[2P(\pi_2)P_{12}+P(\pi_2)P_{32} \\
&\quad +P(\pi_3)P_{13}-\alpha_1] \\
&\quad -\lambda_2[2P(\pi_2)P_{22}-P(\pi_1)P_{31} \\
&\quad +P(\pi_3)P_{23}-\alpha_2], \qquad (54)
\end{aligned}$$

where λ_1 and λ_2 are the Lagrange multipliers.

The functional in (54) is maximized in App. D. The boundary lines which partition the $(\mathbf{LR}_1,\mathbf{LR}_2)$ decision variable plane into the regions Z_1, Z_2, and Z_3 are found to be

$$\begin{aligned}
&\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1-(2\lambda_1-2\lambda_2+1)\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 \\
&\qquad =(\lambda_1-\lambda_2) \qquad (55)
\end{aligned}$$

$$(1-\lambda_2)\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1-\lambda_1\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \lambda_1 \qquad (56)$$

$$-\lambda_2\frac{P(\pi_1)}{P(\pi_3)}\mathrm{LR}_1+(\lambda_1-2\lambda_2+1)\frac{P(\pi_2)}{P(\pi_3)}\mathrm{LR}_2 = \lambda_2 \qquad (57)$$

If we define the quantities $\gamma_{313}\equiv\gamma_{121}[P(\pi_3)/P(\pi_1)]\lambda_1$ and $\gamma_{323}\equiv\gamma_{121}[P(\pi_3)/P(\pi_1)]\lambda_2$, and further require λ_1 and λ_2 to be positive, then the resulting decision strategy is found to be

$$\begin{aligned}
&\gamma_{121}\mathrm{LR}_1-\left[\frac{2\gamma_{313}}{P(\pi_3)}-\frac{2\gamma_{323}}{P(\pi_3)}+\frac{\gamma_{121}}{P(\pi_1)}\right]P(\pi_2)\mathrm{LR}_2 \\
&\qquad =\gamma_{313}-\gamma_{323} \qquad (58)
\end{aligned}$$

$$\begin{aligned}
&\left[\gamma_{121}-\frac{P(\pi_1)}{P(\pi_3)}\gamma_{323}\right]\mathrm{LR}_1-\frac{P(\pi_2)}{P(\pi_3)}\gamma_{313}\mathrm{LR}_2 \\
&\qquad =\gamma_{313} \qquad (59)
\end{aligned}$$

$$\begin{aligned}
&-\frac{P(\pi_1)}{P(\pi_3)}\gamma_{323}\mathrm{LR}_1+\left[\frac{\gamma_{313}}{P(\pi_3)}-\frac{2\gamma_{323}}{P(\pi_3)}\right. \\
&\qquad \left.+\frac{\gamma_{121}}{P(\pi_1)}\right]P(\pi_2)\mathrm{LR}_2 \\
&\qquad =\gamma_{323}. \qquad (60)
\end{aligned}$$

This is in fact the ideal observer subject to the constraints in (49)–(51); that is, the resulting observer is identical to that stated in (39)–(41). This special case of the observer proposed by Scurfield, which we have shown to be a special case of the ideal observer [13], has a performance that depends only on the quantities $P(\pi_1)P_{11}+P(\pi_2)P_{22}$, $2P(\pi_2)P_{12}+P(\pi_2)P_{32}+P(\pi_3)P_{13}$, and $2P(\pi_2)P_{22}-P(\pi_1)P_{31}+P(\pi_3)P_{23}$ by (53). The observer described above is indeed that which obtains optimal performance with respect to this set of quantities related to the conditional classification rates. By the argument at the end of Sec. III-A, this description of the constrained observer's performance is complete.

IV. Discussion

Given the rapid increase in complexity of the utility constraints and performance evaluation criteria as one proceeds from Secs. III-B to III-E, it is quite possible for the main point of the above analyses to become obscured. That main point is that, for each of a variety of constrained special cases of the three-class ideal observer, the performance of that observer is completely describable, in an expected-utility sense, by only two decision criteria and three quantities related to conditional classification rates. This represents a considerable simplification from the general model, which is known to involve five decision criteria and six conditional classification rates. Furthermore, given the result derived in Sec. III-A, this conclusion can be seen to apply to any set of GPDVs obtained from linear constraints on the ideal observer's decision utilities, and not merely the four special cases considered explicitly here. Put another way, it is relatively straightforward to see that if linear restrictions (*i. e.*, constraints) of the form described in Sec. III-A are placed on the ideal observer decision rule, the performance of the resulting observer will be describable with less than five degrees of freedom (or, in general, less than $N^2 - N - 1$ in an N-class classification task). Moreover, we have shown the converse to be true as well for a wide variety of restricted performance evaluation models: if one chooses to describe observer performance with fewer than six (or, in general, fewer than $N^2 - N$) GPDVs that are linearly related to the conditional classification probabilities, then the observer which optimizes performance with respect to that description is a restricted form of the ideal observer (where the restrictions correspond to linear constraints on the utilities). Again, this follows directly from the proof in Sec. III-A that the expected utility and Neyman-Pearson optimization methods are in fact mathematically equivalent.

It should be immediately acknowledged that such simplified models may ultimately prove to be of limited practical importance. Given an observer known to closely approximate the behavior of the unrestricted ideal observer, or indeed given a human observer, it is difficult to conceive of a pragmatic way to externally constrain the observer's decision utilities to match a particular model such as one of those described above. On the other hand, an algorithmic observer (such as an implementation of a computerized scheme for computer-aided diagnosis) might readily allow such constraints on its decision *rules* to be implemented; however, the assumption that the probability density functions of the decision *variables* generated by the scheme do indeed follow those required by the ideal observer model would generally be unverifiable, given the limited amount of data typically available for training and testing such a scheme.

V. Conclusions

Despite the limitations of constrained or simplified performance evaluation models stated in the preceding section, it remains an acknowledged fact that a fully general extension of ROC analysis to classification tasks with three or more classes has yet to be developed. Although the investigation of constrained and therefore tractable observer performance evaluation models should not be considered an end unto itself, a thorough understanding of such models is almost certain to prove necessary for the development of more general observer models. We believe that demonstrating particular constrained ideal observer models to be complete as well as tractable will be a crucial step toward this understanding.

Appendix A
The Chan et al. Observer

As stated in the material leading up to (3)–(5), observer decisions here are assumed to be made based on statistically variable observational data. Explicitly,

$$P_{ij} \equiv \int_{Z_i} p(\vec{x}|\pi_j)\, d^m\vec{x}, \tag{61}$$

where Z_i is the region for which observations $\vec{\mathbf{x}}$ (of dimension m) are decided to belong to the class labeled π_i ($1 \leq i \leq 3$). The expression for F_{Chan} in (16) can then be written as follows:

$$\begin{aligned}
F_{\text{Chan}} &= 1 - P_{13} - P_{23} - \lambda_{31}P_{31} + \lambda_{31}\alpha_{31} - \lambda_{32}P_{32} \\
&\quad + \lambda_{32}\alpha_{32} \\
&= 1 + \lambda_{31}\alpha_{31} + \lambda_{32}\alpha_{32} - \{P_{13} + P_{23} + \lambda_{31}P_{31} \\
&\quad + \lambda_{32}P_{32}\} \\
&= 1 + \lambda_{31}\alpha_{31} + \lambda_{32}\alpha_{32} - \left\{ \int_{Z_1} p(\vec{x}|\pi_3)\, d^m\vec{x} \right. \\
&\quad + \int_{Z_2} p(\vec{x}|\pi_3)\, d^m\vec{x} \\
&\quad \left. + \int_{Z_3} [\lambda_{31}p(\vec{x}|\pi_1) + \lambda_{32}p(\vec{x}|\pi_2)]\; d^m\vec{x} \right\}.
\end{aligned} \tag{62}$$

F_{Chan} is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in (62)) is minimal. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{aligned}
\text{decide } \pi_1 \text{ iff} \quad p(\vec{x}|\pi_3) &< p(\vec{x}|\pi_3) \\
\text{and} \quad p(\vec{x}|\pi_3) &< \lambda_{31}p(\vec{x}|\pi_1) + \lambda_{32}p(\vec{x}|\pi_2) \qquad (63) \\
\text{decide } \pi_2 \text{ iff} \quad p(\vec{x}|\pi_3) &\leq p(\vec{x}|\pi_3) \\
\text{and} \quad p(\vec{x}|\pi_3) &< \lambda_{31}p(\vec{x}|\pi_1) + \lambda_{32}p(\vec{x}|\pi_2) \qquad (64) \\
\text{decide } \pi_3 \text{ iff} \quad p(\vec{x}|\pi_3) &\geq \lambda_{31}p(\vec{x}|\pi_1) + \lambda_{32}p(\vec{x}|\pi_2) \\
\text{and} \quad p(\vec{x}|\pi_3) &\geq \lambda_{31}p(\vec{x}|\pi_1) + \lambda_{32}p(\vec{x}|\pi_2). \qquad (65)
\end{aligned}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\begin{aligned}
\text{decide } \pi_1 \text{ iff} \quad 0\text{LR}_1 - 0\text{LR}_2 &> 0 \\
\text{and} \quad \lambda_{31}\text{LR}_1 + \lambda_{32}\text{LR}_2 &> 1 \qquad (66) \\
\text{decide } \pi_2 \text{ iff} \quad 0\text{LR}_1 - 0\text{LR}_2 &\leq 0 \\
\text{and} \quad \lambda_{31}\text{LR}_1 + \lambda_{32}\text{LR}_2 &> 1 \qquad (67) \\
\text{decide } \pi_3 \text{ iff} \quad \lambda_{31}\text{LR}_1 + \lambda_{32}\text{LR}_2 &\leq 1 \\
\text{and} \quad \lambda_{31}\text{LR}_1 + \lambda_{32}\text{LR}_2 &\leq 1. \qquad (68)
\end{aligned}$$

(We assume without loss of generality that $p(\vec{x}|\pi_3) > 0$, because the task reduces to a two-class problem for values of $\vec{x}$ such that $p(\vec{x}|\pi_3) = 0$.) The corresponding decision boundary lines are given in (17)–(19).

Appendix B
The He et al. Observer

Using (61), the expression for F_{He} in (24) can be expressed as

$$\begin{aligned} F_{\text{He}} &= 1 - P_{13} - P_{23} + \lambda_{11}(1 - P_{21} - P_{31}) - \lambda_{11}\alpha_{11} \\ &\quad + \lambda_{22}(1 - P_{12} - P_{32}) - \lambda_{22}\alpha_{22} \\ &= 1 - \lambda_{11}\alpha_{11} - \lambda_{22}\alpha_{22} - \{P_{13} + P_{23} \\ &\quad + \lambda_{11}(P_{21} + P_{31}) + \lambda_{22}(P_{12} + P_{32})\} \\ &= 1 - \lambda_{11}\alpha_{11} - \lambda_{22}\alpha_{22} \\ &\quad - \left\{ \int_{Z_1} [\lambda_{22} p(\vec{x}|\pi_2) + p(\vec{x}|\pi_3)]\ d^m\vec{x} \right. \\ &\quad + \int_{Z_2} [\lambda_{11} p(\vec{x}|\pi_1) + p(\vec{x}|\pi_3)]\ d^m\vec{x} \\ &\quad \left. + \int_{Z_3} [\lambda_{11} p(\vec{x}|\pi_1) + \lambda_{22} p(\vec{x}|\pi_2)]\ d^m\vec{x} \right\}. \qquad (69) \end{aligned}$$

F_{He} is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in (69)) is minimal. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{aligned} \text{decide } \pi_1 \text{ iff} \quad \lambda_{22} p(\vec{x}|\pi_2) &< \lambda_{11} p(\vec{x}|\pi_1) \\ \text{and} \quad p(\vec{x}|\pi_3) &< \lambda_{11} p(\vec{x}|\pi_1) \qquad (70) \\ \text{decide } \pi_2 \text{ iff} \quad \lambda_{11} p(\vec{x}|\pi_1) &\leq \lambda_{22} p(\vec{x}|\pi_2) \\ \text{and} \quad p(\vec{x}|\pi_3) &< \lambda_{22} p(\vec{x}|\pi_2) \qquad (71) \\ \text{decide } \pi_3 \text{ iff} \quad \lambda_{11} p(\vec{x}|\pi_1) &\leq p(\vec{x}|\pi_3) \\ \text{and} \quad \lambda_{22} p(\vec{x}|\pi_2) &\leq p(\vec{x}|\pi_3). \qquad (72) \end{aligned}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\begin{aligned} \text{decide } \pi_1 \text{ iff} \quad \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 &> 0 \\ \text{and} \quad \lambda_{11}\text{LR}_1 &> 1 \qquad (73) \\ \text{decide } \pi_2 \text{ iff} \quad \lambda_{11}\text{LR}_1 - \lambda_{22}\text{LR}_2 &\leq 0 \\ \text{and} \quad \lambda_{22}\text{LR}_2 &> 1 \qquad (74) \\ \text{decide } \pi_3 \text{ iff} \quad \lambda_{11}\text{LR}_1 &\leq 1 \\ \text{and} \quad \lambda_{22}\text{LR}_2 &\leq 1. \qquad (75) \end{aligned}$$

The corresponding decision boundary lines are given in (25)–(27).

Appendix C
The Scurfield Observer (Likelihood Ratio)

Using (61), the expression for $F_{\text{Scfd:LR}}$ in (35) can be written as

$$\begin{aligned} F_{\text{Scfd:LR}} &= 1 - P_{21} - P_{31} + 1 - P_{12} - P_{32} - \lambda_{13}P_{13} \\ &\quad + \lambda_{13}\alpha_{13} - \lambda_{23}P_{23} + \lambda_{23}\alpha_{23} \\ &= 2 + \lambda_{13}\alpha_{13} + \lambda_{23}\alpha_{23} - \{P_{21} + P_{31} + P_{12} \\ &\quad + P_{32} + \lambda_{13}P_{13} + \lambda_{23}P_{23}\} \\ &= 2 + \lambda_{13}\alpha_{13} + \lambda_{23}\alpha_{23} \\ &\quad - \left\{ \int_{Z_1} [p(\vec{x}|\pi_2) + \lambda_{13} p(\vec{x}|\pi_3)]\ d^m\vec{x} \right. \\ &\quad + \int_{Z_2} [p(\vec{x}|\pi_1) + \lambda_{23} p(\vec{x}|\pi_3)]\ d^m\vec{x} \\ &\quad \left. + \int_{Z_3} [p(\vec{x}|\pi_1) + p(\vec{x}|\pi_2)]\ d^m\vec{x} \right\}. \qquad (76) \end{aligned}$$

$F_{\text{Scfd:LR}}$ is maximized when the quantity in braces is minimized. This quantity, in turn, can be minimized by assigning a given $\vec{x}$ to the region Z_i such that the ith integrand (from among the integrals in braces in (76)) is minimal. (Situations in which two or more of the integrands yield the same minimal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{aligned} &\text{decide } \pi_1 \text{ iff} \\ &\qquad p(\vec{x}|\pi_2) + \lambda_{13} p(\vec{x}|\pi_3) < p(\vec{x}|\pi_1) + \lambda_{23} p(\vec{x}|\pi_3) \\ &\quad \text{and} \qquad \lambda_{13} p(\vec{x}|\pi_3) < p(\vec{x}|\pi_1) \qquad (77) \\ &\text{decide } \pi_2 \text{ iff} \\ &\qquad p(\vec{x}|\pi_1) + \lambda_{23} p(\vec{x}|\pi_3) \leq p(\vec{x}|\pi_2) + \lambda_{13} p(\vec{x}|\pi_3) \\ &\quad \text{and} \qquad \lambda_{23} p(\vec{x}|\pi_3) < p(\vec{x}|\pi_2) \qquad (78) \\ &\text{decide } \pi_3 \text{ iff} \\ &\qquad p(\vec{x}|\pi_1) \leq \lambda_{13} p(\vec{x}|\pi_3) \\ &\quad \text{and} \qquad p(\vec{x}|\pi_2) \leq \lambda_{23} p(\vec{x}|\pi_3). \qquad (79) \end{aligned}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ to obtain

$$\begin{aligned} \text{decide } \pi_1 \text{ iff} \quad \text{LR}_1 - \text{LR}_2 &> \lambda_{13} - \lambda_{23} \\ \text{and} \quad \text{LR}_1 &> \lambda_{13} \qquad (80) \\ \text{decide } \pi_2 \text{ iff} \quad \text{LR}_1 - \text{LR}_2 &\leq \lambda_{13} - \lambda_{23} \\ \text{and} \quad \text{LR}_2 &> \lambda_{23} \qquad (81) \\ \text{decide } \pi_3 \text{ iff} \quad \text{LR}_1 &\leq \lambda_{13} \\ \text{and} \quad \text{LR}_2 &\leq \lambda_{23}. \qquad (82) \end{aligned}$$

The corresponding decision boundary lines are given in (36)–(38).

Appendix D
The Scurfield Observer (a posteriori Class Probability)

Using (61), the expression for $F_{\text{Scfd:AP}}$ in (54) can be written as

$$\begin{aligned} F_{\text{Scfd:AP}} &= \lambda_1\alpha_1 + \lambda_2\alpha_2 \\ &\quad + P(\pi_1) \int_{Z_1} p(\vec{x}|\pi_1)\, d^m\vec{x} + P(\pi_2) \int_{Z_2} p(\vec{x}|\pi_2)\, d^m\vec{x} \\ &\quad - \lambda_1 \left[2P(\pi_2) \int_{Z_1} p(\vec{x}|\pi_2)\, d^m\vec{x} \right. \\ &\quad \left. + P(\pi_2) \int_{Z_3} p(\vec{x}|\pi_2)\, d^m\vec{x} + P(\pi_3) \int_{Z_1} p(\vec{x}|\pi_3)\, d^m\vec{x} \right] \\ &\quad - \lambda_2 \left[2P(\pi_2) \int_{Z_2} p(\vec{x}|\pi_2)\, d^m\vec{x} \right. \\ &\quad - P(\pi_1) \int_{Z_3} p(\vec{x}|\pi_1)\, d^m\vec{x} \end{aligned}$$

$$+ P(\pi_3) \int_{Z_2} p(\vec{x}|\pi_3)\, d^m\vec{x} \Bigg] . \tag{83}$$

Collecting terms with given domains of integration yields

$$\begin{aligned} F_{\text{Scfd:AP}} &= \lambda_1\alpha_1 + \lambda_2\alpha_2 \\ &\quad + \int_{Z_1} [P(\pi_1)p(\vec{x}|\pi_1) - 2\lambda_1 P(\pi_2)p(\vec{x}|\pi_2) \\ &\quad - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3)]\ d^m\vec{x} \\ &\quad + \int_{Z_2} [P(\pi_2)p(\vec{x}|\pi_2) - 2\lambda_2 P(\pi_2)p(\vec{x}|\pi_2) \\ &\quad - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3)]\ d^m\vec{x} \\ &\quad + \int_{Z_3} [-\lambda_1 P(\pi_2)p(\vec{x}|\pi_2) \\ &\quad + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1)]\ d^m\vec{x}. \end{aligned} \tag{84}$$

$F_{\text{Scfd:AP}}$ can be maximized by assigning a given $\vec{x}$ to the region Z_i such that the integrand over Z_i in (84) is maximal. (Situations in which two or more of the integrands yield the same maximal value for a given $\vec{x}$ can be decided in an arbitrary but consistent fashion.)

That is,

$$\begin{aligned} &\text{decide } \pi_1 \text{ iff} \\ &P(\pi_1)p(\vec{x}|\pi_1) - 2\lambda_1 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\ &\quad > P(\pi_2)p(\vec{x}|\pi_2) - 2\lambda_2 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\ &\text{and} \quad P(\pi_1)p(\vec{x}|\pi_1) - 2\lambda_1 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\ &\quad > \quad -\lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \end{aligned} \tag{85}$$

$$\begin{aligned} &\text{decide } \pi_2 \text{ iff} \\ &P(\pi_2)p(\vec{x}|\pi_2) - 2\lambda_2 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\ &\quad \geq P(\pi_1)p(\vec{x}|\pi_1) - 2\lambda_1 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\ &\text{and} \quad P(\pi_2)p(\vec{x}|\pi_2) - 2\lambda_2 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\ &\quad > \quad - \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \end{aligned} \tag{86}$$

$$\begin{aligned} &\text{decide } \pi_3 \text{ iff} \\ &P(\pi_1)p(\vec{x}|\pi_1) - 2\lambda_1 P(\pi_2)p(\vec{x}|\pi_3) - \lambda_1 P(\pi_3)p(\vec{x}|\pi_3) \\ &\quad \leq - \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1) \\ &\text{and} \quad P(\pi_2)p(\vec{x}|\pi_2) - 2\lambda_2 P(\pi_2)p(\vec{x}|\pi_2) - \lambda_2 P(\pi_3)p(\vec{x}|\pi_3) \\ &\quad \leq \quad - \lambda_1 P(\pi_2)p(\vec{x}|\pi_2) + \lambda_2 P(\pi_1)p(\vec{x}|\pi_1). \end{aligned} \tag{87}$$

We can divide these relations by $p(\vec{x}|\pi_3)$ and rearrange terms to obtain

$$\begin{aligned} \text{decide } \pi_1 \text{ iff} \quad & P(\pi_1)\mathrm{LR}_1 - (2\lambda_1 - 2\lambda_2 + 1)P(\pi_2)\mathrm{LR}_2 \\ & > (\lambda_1 - \lambda_2)P(\pi_3) \\ \text{and} \quad & (1 - \lambda_2)P(\pi_1)\mathrm{LR}_1 - \lambda_1 P(\pi_2)\mathrm{LR}_2 \\ & > \lambda_1 P(\pi_3) \end{aligned} \tag{88}$$

$$\begin{aligned} \text{decide } \pi_2 \text{ iff} \quad & P(\pi_1)\mathrm{LR}_1 - (2\lambda_1 - 2\lambda_2 + 1)P(\pi_2)\mathrm{LR}_2 \\ & \leq (\lambda_1 - \lambda_2)P(\pi_3) \\ \text{and} \quad & -\lambda_2 P(\pi_1)\mathrm{LR}_1 + (\lambda_1 - 2\lambda_2 + 1)P(\pi_2)\mathrm{LR}_2 \\ & > \lambda_2 P(\pi_3) \end{aligned} \tag{89}$$

$$\begin{aligned} \text{decide } \pi_3 \text{ iff} \quad & (1 - \lambda_2)P(\pi_1)\mathrm{LR}_1 - \lambda_1 P(\pi_2)\mathrm{LR}_2 \\ & \leq \lambda_1 P(\pi_3) \\ \text{and} \quad & -\lambda_2 P(\pi_1)\mathrm{LR}_1 + (\lambda_1 - 2\lambda_2 + 1)P(\pi_2)\mathrm{LR}_2 \\ & \leq \lambda_2 P(\pi_3). \end{aligned} \tag{90}$$

The corresponding decision boundary lines are given in (55)–(57).

Acknowledgment

The authors would like to thank the associate editor and reviewers for their insightful comments which greatly helped to improve both the structure and content of this paper.

References

[1] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.

[2] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. VIII, no. 4, pp. 283–298, 1978.

[3] D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.*, vol. 31, pp. 81–90, 2004.

[4] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I.* New York: John Wiley & Sons, 1968.

[5] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.*, vol. 23, pp. 891–895, 2004.

[6] C. E. Metz and X. Pan, " 'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.*, vol. 43, pp. 1–33, 1999.

[7] D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks," *IEEE Trans. Med. Imag.*, vol. 24, pp. 293–299, 2005.

[8] D. C. Edwards and C. E. Metz, "Restrictions on the three-class ideal observer's decision boundary lines," *IEEE Trans. Med. Imag.*, vol. 24, pp. 1566–1573, 2005.

[9] B. K. Scurfield, "Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks," *J. Math. Psychol.*, vol. 42, pp. 5–31, 1998.

[10] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, pp. 78–89, 1999.

[11] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study," in Proc. SPIE Vol. 5032 *Medical Imaging 2003: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, Eds., SPIE, Bellingham, WA, 2003, pp. 567–578.

[12] X. He, C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis — A decision theoretic approach under the ideal observer framework," *IEEE Trans. Med. Imag.*, vol. 25, pp. 571–581, 2006.

[13] D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.*, vol. 50, pp. 478–487, 2006.

[14] S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations: Second Edition.* San Diego, CA: Harcourt Brace Jovanovich, 1986.

J A utility-based performance metric for ROC analysis of N-class classification tasks

A utility-based performance metric for ROC analysis of N-class classification tasks

Darrin C. Edwards* and Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

We have shown previously that an obvious generalization of the area under an ROC curve (AUC) cannot serve as a useful performance metric in classification tasks with more than two classes. We define a new performance metric, grounded in the concept of expected utility familiar from ideal observer decision theory, but which should not suffer from the issues of dimensionality and degeneracy inherent in the hypervolume under the ROC hypersurface in tasks with more than two classes. In the present work, we compare this performance metric with the traditional AUC metric in a variety of two-class tasks. Our numerical studies suggest that the behavior of the proposed performance metric is consistent with that of the AUC performance metric in a wide range of two-class classification tasks, while analytical investigation of three-class "near-guessing" observers supports our claim that the proposed performance metric is well-defined and positive in the limit as the observer's performance approaches that of the guessing observer.

Keywords: ROC methodology, expected utility, three-class classification

1. INTRODUCTION

We are attempting to extend the well-known observer performance evaluation methodology of receiver operating characteristic (ROC) analysis[1, 2] to classification tasks with three or more classes. This could conceivably be of benefit, for example, in a medical decision-making task in which a region of a patient image must be characterized as containing a malignant lesion, a benign lesion, or only normal tissue.[3]

Unfortunately, a fully general but tractable extension of ROC analysis to tasks with more than two classes has yet to be developed. It is known that the performance of an observer in a classification task with N classes ($N \geq 2$) can be completely described by a set of $N^2 - N$ conditional error probabilities,[4, 5] and that the performance of the ideal observer (that which minimizes Bayes risk[4]) is completely characterized by an ROC hypersurface in which these conditional error probabilities depend on a set of $N^2 - N - 1$ decision criteria.[5] Although analytic expressions for the ideal observer's conditional error probabilities given reasonable models for the underlying observational date have been worked out in the two-class case,[6] this has not yet been accomplished in a fully general manner for tasks with three or more classes.

Furthermore, we have shown that an obvious generalization of the area under the ROC curve (AUC) does not in fact yield a useful performance metric in tasks with three or more classes.[7] In the formulation we advocate, the set of $N^2 - N$ conditional error probabilities serve as the axes of the observer's ROC space. This is equivalent to plotting a two-class observer's false-negative fraction (FNF), rather than the more conventional true-positive fraction (TPF), as a function of false-positive fraction (FPF) to construct the observer's ROC curve. Since $\mathrm{FNF} = 1 - \mathrm{TPF}$, this yields an ROC curve which is simply an "upside-down" version of the conventional curve, and the area under this ROC curve (which we will denote $\widetilde{A}$) is just one minus the conventionally defined AUC. Clearly this area will vary from 0.5, for a "guessing" observer, to 0, for a "perfect" observer. In a task with more than two classes, however, we showed that although the "hypervolume under the ROC hypersurface" (HUH) is again 0 for a perfect observer, the HUH of a guessing observer is, counterintuitively, also 0.[7] (Briefly, the number of degrees of freedom of the guessing observer's ROC hypersurface is $N - 1$ rather than $N^2 - N - 1$, yielding a "degenerate" hypersurface with no hypervolume, much as in three dimensions the integral under a "surface" which is actually a curve — *e.g.*, $z = f(x, y)$ where $y = g(x)$ — will be zero.)

*Correspondence: E-mail: d-edwards@uchicago.edu; Telephone: 773 834 5094; Fax: 773 702 0371

What is needed is a performance metric that shares the useful properties of AUC, namely its intuitive direct relationship to the "difficulty" of the observer's task ("near-guessing" observers have an $\widetilde{A}$ near 0.5, "near-perfect" observers have an $\widetilde{A}$ near 0), without suffering from this drawback of degeneracy. We have begun to investigate a performance metric that has its origins in the "expected utility" concept fundamental to ideal observer decision theory,[4] and which we have reason to believe is both related to HUH and yet not plagued by the degeneracy issues of the HUH. In the next section, we attempt to motivate this performance metric, the "surface-averaged expected cost" (SAEC), and derive theoretical properties of this quantity. In Sec. 3, we outline the simulation studies we implemented in a number of simple two-class classification tasks; the results of those studies are presented in Sec. 4. The implications and limitations of the proposed metric are discussed in Sec. 5, and we summarize our conclusions in Sec. 6.

2. THEORY

In a two-class classification task, with the classes labeled "π_+" ("positive") and "π_-" ("negative"), the expected utility of an observer can be written as[4]

$$E\{\mathbf{U}\} \equiv (U_{TP}\mathrm{TPF} + U_{FN}\mathrm{FNF})P(\pi_+) + (U_{FP}\mathrm{FPF} + U_{TN}\mathrm{TNF})P(\pi_-), \tag{1}$$

where TPF is the probability of deciding an observation is positive, conditional on it actually being drawn from class π_+, more explicitly denoted as $P(\mathbf{d} = \pi_+|\mathbf{t} = \pi_+)$; FNF is $P(\mathbf{d} = \pi_-|\mathbf{t} = \pi_+)$; FPF is $P(\mathbf{d} = \pi_+|\mathbf{t} = \pi_-)$; and TNF is the true-negative fraction, or $P(\mathbf{d} = \pi_-|\mathbf{t} = \pi_-)$. Each U represents the utility of a particular decision under a particular truth condition. (We use a bold typeface to denote statistically variable quantities, and here $\mathbf{t}$ denotes the true class to which a randomly sampled observation belongs, while $\mathbf{d}$ denotes the decision made for that observation.)

In a classification task with an arbitrary number of classes N, with labels running from π_1 to π_N, the above expression is readily generalized to obtain

$$E\{\mathbf{U}\} \equiv \sum_{j=1}^{N}\sum_{i=1}^{N}(U_{i|j}P_{ij})P(\pi_j), \tag{2}$$

where we have written the observer's conditional classification rates $P(\mathbf{d} = \pi_i|\mathbf{t} = \pi_j)$ simply as P_{ij}. From the rules for conditional probability,[8] $\sum_i P_{ij} = 1$, and so we can rewrite this expression to obtain

$$\begin{aligned} E\{\mathbf{U}\} &= \sum_{i=1}^{N} U_{i|i}P(\pi_i) \\ &\quad -\sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}(U_{j|j} - U_{i|j})P(\pi_j)P_{ij} \\ &= U_0 - \sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\gamma_{jij}P_{ij}, \end{aligned} \tag{3}$$

where U_0 is just the expression $\sum_i U_{i|i}P(\pi_i)$ (independent of the conditional error rates P_{ij} which describe the observer's performance), and $\gamma_{jij} \equiv (U_{j|j} - U_{i|j})P(\pi_j)$ gives, to within an arbitrary scale factor, the set of $N^2 - N - 1$ decision criteria used by the ideal observer to make decisions.[5,9–11] Note that the γ_{jij} are strictly positive if we impose the reasonable assumption that an incorrect utility will always have a smaller utility than the corresponding correct decision. If we now define the "normalized" utility (more precisely, if we choose particular units in which to "measure" utility) as

$$\mathbf{u} \equiv \frac{\mathbf{U}}{(\sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\gamma_{jij}^2)^{1/2}}, \tag{4}$$

and similarly define $\gamma_0 \equiv U_0/(\sum \gamma_{jij}^2)^{1/2}$, we can simplify the expression for expected utility further to obtain

$$E\{\mathbf{u}\} = \gamma_0 - \hat{\gamma} \cdot \vec{P}. \tag{5}$$

Here $\vec{P}$ is an $(N^2 - N)$-dimensional vector whose components are the conditional error rates P_{ij} (with a specified ordering, *e.g.*, $(P_{12}, P_{13}, \ldots, P_{1N}, P_{21}, \ldots, P_{N(N-2)}, P_{N(N-1)})$) — *i.e.*, the coordinates of ROC space; and $\hat{\gamma}$ is a unit vector of the same dimensionality as $\vec{P}$, whose components are the corresponding values of γ_{jij} after normalization.

It is important to keep in mind that although this normalized expected utility is optimized only by the ideal observer, it is well-defined for any observer at a particular operating point $\vec{P}$ and choice of (normalized) utilities *via* $\hat{\gamma}$. Furthermore, assuming the values of the observational priors $P(\pi_i)$ to be fixed and the values of the utilities to be determined externally to the observer (*i.e.*, not modifiable by the observer within a given experiment or set of experiments), maximizing the normalized expected utility is clearly equivalent to minimizing $\hat{\gamma} \cdot \vec{P}$. We will refer to this latter quantity as the expected cost; note that although "cost" has a far more general definition in the literature (as do "utility," "risk," etc.), we will attempt to avoid confusion here by using the term only in this restricted sense.

Suppose we have measured the set of all possible values of $P_{N(N-1)}$ for a given observer as a function of the other $N^2 - N - 1$ conditional error probabilities. (For the ideal observer, this can be conceived of as measuring $\vec{P}$ for every possible value of $\hat{\gamma}$; for a non-ideal observer, we assume that we can modify whatever set of $N^2 - N - 1$ decision criteria it is actually using, even if these are not usefully related to the utilities.) We write this as

$$\begin{aligned} P_{N(N-1)} &= R(P_{12}, P_{13}, \ldots, P_{1N}, P_{21}, \ldots, P_{N(N-3)}, P_{N(N-2)}) \\ &= R(\vec{P}^*), \end{aligned} \tag{6}$$

where $\vec{P}^*$ denotes the "reduced" vector, of dimensionality $N^2 - N - 1$, obtained by deleting the $(N^2 - N)$th component of $\vec{P}$. The HUH can be defined[7] as

$$\text{HUH} \equiv \int\limits_{\Omega_R} R(\vec{P}^*)\, d^{N^2-N-1}\vec{P}^*, \tag{7}$$

or equivalently,

$$\text{HUH} \equiv \int\limits_{V_R} d^{N^2-N}\vec{P}, \tag{8}$$

where Ω_R denotes the set of $\vec{P}^*$ for which $R(\vec{P}^*)$ is defined (the domain of the function defining the ROC hypersurface), and V_R denotes the set of all $\vec{P}$ enclosed by that hypersurface and by the boundaries of the ROC space (given that $0 \le P_{ij} \le 1$). Note that in a two-class task, with the ROC curve given by FNF $= R$(FPF), this reduces to

$$\begin{aligned} \text{HUH} &= \int\limits_{V_R} d^{N^2-N}\vec{P} \\ &= \int\limits_0^1 \int\limits_0^{R(\text{FPF})} d\text{FNF}\, d\text{FPF} \\ &= \int\limits_0^1 R(\text{FPF})\, d\text{FPF} \\ &= \widetilde{A}, \end{aligned} \tag{9}$$

as expected. (Note that, as stated in Sec. 1, this is one minus the conventional AUC that would be obtained by integrating TPF as a function of FPF.)

Despite the long-standing success of AUC as a summary performance metric for ROC analysis, we have shown the HUH not to be useful for this purpose in a classification task with three or more classes.[7] Briefly, a "perfect" observer can achieve values of, say, $P_{N(N-1)} = 0$ for any achievable set of $\vec{P}^*$; by Eq. 7, the HUH for such an observer will thus be zero (and will approach zero for a "near-perfect" observer). A "guessing" observer will assign observations to the N classes randomly, independent of the actual truth states of those observations; since the total probability of making a decision will be one, this leaves a set of only $N-1$ degrees of freedom (each of the probabilities of assigning an observation to a given class). But it can be shown that in such a situation, the resulting domain of integration Ω_R is "degenerate," and the integral in Eq. 7 is zero regardless of the value of the integrand (and will approach zero for a "near-guessing" observer). Thus, opposite extremes of performance result in similar or identical values of HUH, making this quantity useless even as a summary performance metric in classification tasks with more than two classes. In the two-class case, $N^2 - N - 1 = N - 1 = 1$, of course, and (amusingly or providentially, depending perhaps on one's worldview) no such degeneracy is encountered.

Discouraging though this result may be, it immediately brings to the forefront the question of what motivated the choice of AUC as a summary performance metric to begin with. In the present context, it can be said that AUC averages directly over "performance description variables" (such as FNF) without regard to utility (or, equivalently, cost). For an experiment involving a human observer (the internals of whose decision-making process may be unavailable to experimenter control) or an algorithmic observer (trained on a finite sample of observational data), the actual "costs" may be unknown to the experimenter, or may not be available for modification in any practical sense. On the other hand, ideal observer decision theory demonstrates the tremendous theoretical and practical importance of Eq. 5, and it is natural to ask whether consideration of the expected cost, $\hat{\gamma} \cdot \vec{P}$, might not be worthwhile, given the difficulty in generalizing AUC just described.

For the ideal observer itself, this line of inquiry seems quite promising indeed. For each possible value of $\hat{\gamma}$, the ideal observer will choose an operating point $\vec{P}$ that minimizes the expected cost. (It is possible, given particular forms of the observational data probability density functions (PDFs), that multiple operating points $\vec{P}$ will be associated with a given $\hat{\gamma}$; it can be shown, however, that such points will always lie in a simply connected region, analogous to a straight line along a two-class ROC surface. We will not consider such special cases here.) By taking the ideal observer's ROC hypersurface as given, one can proceed in the opposite direction: at any given point on the ideal observer's ROC surface, the appropriate $\hat{\gamma}$ for that point is that which minimizes the expected cost. This, in turn, can be shown to imply that the appropriate $\hat{\gamma}$ is normal to the ideal observer's ROC hypersurface at each point $\vec{P}$.

For non-ideal observers, the situation is much more confusing. Given that such an observer might not be basing its decisions on the utilities (available to the ideal observer) at all, it is unclear what value of $\hat{\gamma}$ to assign to a given $\vec{P}$ on such an observer's ROC surface. Arbitrarily, we choose to make the same assignment made by the ideal observer: at each point on the observer's ROC hypersurface, we choose that value of $\hat{\gamma}$ that is normal to the ROC hypersurface at that point. Intuitively, this can be taken to be giving the non-ideal observer the "benefit of the doubt": in determining a total expected cost for the observer, we will at each point take the contribution to that cost to be the "minimum" possible. Alternatively, we can say that the observer under this model is at least behaving "locally" optimally.

Thus, for the ROC hypersurface given in Eq. 6, we define the "local" utility vector to be

$$\hat{\gamma}_R \equiv \frac{(-\nabla R, 1)}{\sqrt{|\nabla R|^2 + 1}}, \tag{10}$$

where the expression in parentheses denotes a vector of dimension $N^2 - N$ whose first $N^2 - N - 1$ components are the negatives of the components of ∇R; the sign is chosen because the components of $\hat{\gamma}_R$ must be positive, ruling out the possibility $(\nabla R, -1)$. We use this definition to construct the surface integral

$$\int_{\sigma_R} \hat{\gamma}_R \cdot \vec{P} \; d^{N^2-N-1}\sigma. \tag{11}$$

The integral is over the ROC hypersurface σ_R, that is, the set of points $\vec{P}$ such that $P_{N(N-1)} = R(\vec{P}^*)$.The differential element on this hypersurface is denoted by $d^{N^2-N-1}\sigma$, where the superscript reminds us of the dimensionality "within" that surface.

In the two class case, the differential element reduces to the differential arc length, which we can define as

$$ds \equiv \sqrt{1+\left(\frac{d\mathrm{FNF}}{d\mathrm{FPF}}\right)^2}\, d\mathrm{FPF}. \tag{12}$$

The integral in Eq. 11 can then be written as

$$\begin{aligned}
\int_0^1 \frac{(\frac{-d\mathrm{FNF}}{d\mathrm{FPF}},1)}{\sqrt{\left(\frac{d\mathrm{FNF}}{d\mathrm{FPF}}\right)^2+1}} \cdot (FPF,FNF)\sqrt{1+\left(\frac{d\mathrm{FNF}}{d\mathrm{FPF}}\right)^2}\, d\mathrm{FPF} &= \int_0^1 \left(-\mathrm{FPF}\frac{d\mathrm{FNF}}{d\mathrm{FPF}}+\mathrm{FNF}\right) d\mathrm{FPF} \\
&= \int_0^1 -\mathrm{FPF}\frac{d\mathrm{FNF}}{d\mathrm{FPF}}\, d\mathrm{FPF} + \int_0^1 \mathrm{FNF}\, d\mathrm{FPF} \\
&= \int_0^1 \mathrm{FPF}\, d\mathrm{FNF} + \int_0^1 \mathrm{FNF}\, d\mathrm{FPF} \\
&= 2\widetilde{A}.
\end{aligned} \tag{13}$$

Note that in the next to last step, the negative sign has disappeared because FNF $= 0$ when FPF $= 1$ and vice versa, so that the order of the limits of integration will be reversed. It is also vital to remember that $\widetilde{A}$ here denotes the area under the "upside-down" ROC curve (FNF plotted against FPF), and is thus one minus the conventional AUC.

Clearly the quantity we have defined is directly related to performance — in fact, far more closely than we had reason to hope: despite our *ad hoc* choice of $\hat{\gamma}_R$, the relation in Eq. 13 holds for arbitrary observers, and not just ideal observers. Even more surprisingly, the generalization of this relationship can be shown to hold for observers in tasks with arbitrary numbers of classes. Returning to Eq. 11, we rearrange terms to obtain

$$\begin{aligned}
\int_{\sigma_R} \hat{\gamma}_R \cdot \vec{P}\, d^{N^2-N-1}\sigma &= \int_{\partial V_R} \hat{\gamma}_R \cdot \vec{P}\, d^{N^2-N-1}\sigma \\
&= \int_{\partial V_R} \vec{P} \cdot \hat{\gamma}_R\, d^{N^2-N-1}\sigma \\
&= \int_{\partial V_R} \vec{P} \cdot \left[\frac{(-\nabla R, 1)}{\sqrt{|\nabla R|^2+1}}\, d^{N^2-N-1}\sigma\right] \\
&= \int_{V_R} \mathrm{div}\vec{P}\, d^{N^2-N}\vec{P} \\
&= (N^2-N)\mathrm{HUH}.
\end{aligned} \tag{14}$$

Here we have used the n-dimensional extension of the divergence theorem (known in three dimensions as Gauss's theorem);[12] div is the operator $\sum_i(\partial/\partial P_i)$, which when applied to the vector $\vec{P}$ will simply yield the dimensionality N^2-N of $\vec{P}$. Note also that in the first step, we have "closed" the ROC hypersurface with the boundary ∂V_R of the ROC hypervolume; this can be done for the given integrand, because the "bottom" surface $P_{N(N-1)} = 0$ will contribute nothing to the surface integral.

Unfortunately, we are now back where we started: since it is equal (to within a proportionality constant) to the HUH, the surface integral defined above will have exactly the same drawbacks as that quantity. However, writing the performance metric in this form — as an integral of the scalar quantity $\hat{\gamma}_R \cdot \vec{P}$ over the ROC hypersurface — suggests a different approach, namely, considering an "average" of this quantity over the hypersurface:

$$\overline{C}_\sigma \equiv \frac{\int\limits_{\sigma_R} \hat{\gamma}_R \cdot \vec{P} \; d^{N^2-N-1}\sigma}{\int\limits_{\sigma_R} d^{N^2-N-1}\sigma} \tag{15}$$

where we have divided the previous quantity by the "surface area" of the ROC hypersurface. The quantity $\overline{C}_\sigma$ is the SAEC referred to in Sec. 1; the overline reminds us that it is an expectation value, and the subscript σ reminds us that it is averaged over a surface (the ROC hypersurface). This can be considered analogous to the concept from univariable calculus of the "average" of a function over an interval:

$$f\text{avg} \equiv \frac{1}{b-a} \int\limits_a^b f(x) \; dx. \tag{16}$$

In particular, it should be immediately clear that $\overline{C}_\sigma$ is bounded by the maximum and minimum values of $\hat{\gamma}_R \cdot \vec{P}$, and that if $\hat{\gamma}_R \cdot \vec{P}$ were constant over a given ROC hypersurface, then $\overline{C}_\sigma$ would be equal to this constant value.

Further analysis will need to be performed to confirm that this quantity remains well-defined for guessing or even "near-guessing" observers. We have reason to believe that an extension of L'Hôpital's rule should be applicable in this case; *i.e.*, although the numerator and denominator will both converge to zero in the limit of approach to a guessing observer, the limit of $\overline{C}_\sigma$ itself should still be a non-zero quantity. Our results in this regard, however, are still very preliminary. For the present work, we will consider only properties of this quantity in the two-class case, where the degeneracy issues involving HUH do not arise. In the two-class case, of course, we can use Eq. 13 to write

$$\overline{C}_\sigma \equiv \frac{2\widetilde{A}}{S} \tag{17}$$

where S is the arc-length along the ROC curve.

3. MATERIALS AND METHOD

We numerically investigated the behavior of $\overline{C}_\sigma$ compared with the conventional AUC under two models for the distributions of the observer's latent decision variable data: the "conventional" binormal model,[13] and the ideal-observer-related "proper" binormal model.[6] Under the conventional model, the observer's decision variables are assumed to be drawn from a pair of distributions which are an (unspecified) monotonic transformation of two normal distributions:

$$\begin{aligned} \mathbf{x}_+ \quad &\sim \quad N(x; \mu_+ = a/b, \sigma_+ = 1/b) \qquad (18) \\ &\text{and} \\ \mathbf{x}_- \quad &\sim \quad N(x; \mu_- = 0, \sigma_- = 1), \qquad (19) \end{aligned}$$

where $N(x; \mu, \sigma)$ is a normal density function with mean μ and standard deviation σ. The observer makes decisions by comparing an observation of unknown class $\mathbf{x}$ with a threshold x_0; varying this threshold from $-\infty$ to ∞ will sweep out the observer's ROC curve. This curve is completely specified by the two parameters a and b, and analytic forms exist for both individual operating points (FPF, TPF) and the conventional AUC (denoted A_z under this model) as functions of a and b.[13]

Under the "proper" binormal model, the observer is again assumed to make decisions using underlying data monotonically related to the pair of distributions given in Eqs. 18 and 19. However, the actual decisions are made by comparing the likelihood ratio of $\mathbf{x}$, rather than $\mathbf{x}$ itself, with a threshold. The likelihood ratio is given by

$$\mathbf{y} \equiv \frac{N(\mathbf{x}; a/b, 1/b)}{N(\mathbf{x}; 0, 1)}. \tag{20}$$

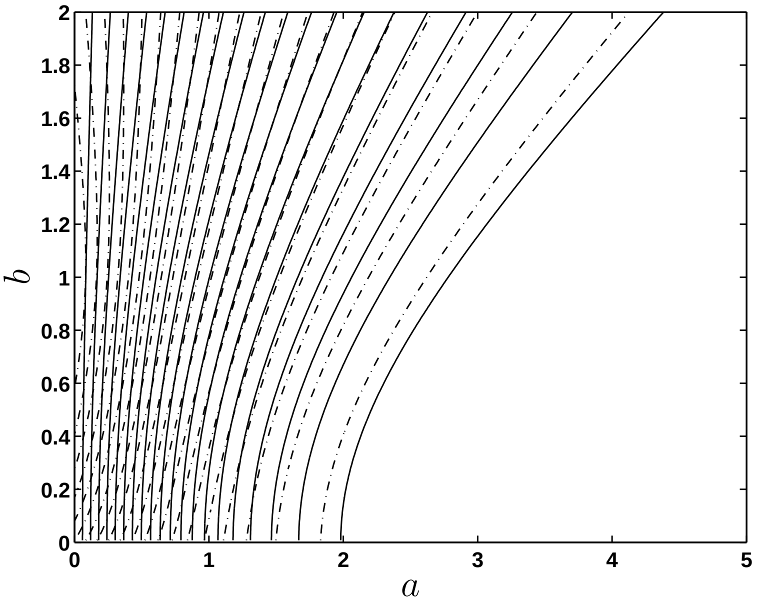


Figure 1. Isopleths of the A_z performance metric (solid lines) and of the proposed $\overline{C}_\sigma$ metric (dash-dotted lines), for various values of the a and b parameters of the conventional binormal model.

Varying the threshold y_0 throughout its range will sweep out the observer's ROC curve. For numerical purposes, it has been found convenient to parametrize this curve using the parameters

$$c \equiv \frac{b-1}{b+1} \tag{21}$$

and

$$d_a \equiv \frac{\sqrt{2}a}{\sqrt{1+b^2}} \tag{22}$$

rather than a and b directly. The observer's ROC curve is completely specified by c and d_a, and analytic forms have been determined for both individual operating points $(\mathrm{FPF}, \mathrm{TPF})$ and the conventional AUC under this model as functions of those two parameters.[6]

We calculated the A_z of an observer assumed to operate under the conventional binormal model for 250 values of a distributed uniformly between 0 and 5, and (at each such value of a) for 250 values of b distributed uniformly between 0 and 2. For each of these 62,500 pairs of parameter values, we also calculated the corresponding value of $\overline{C}_\sigma$ using the relation in Eq. 17. (The arc length S was calculated by generating a large number of operating points along the curve, and adding together the line segment lengths $\sqrt{(FPF_i - FPF_{i-1})^2 + (TPF_i - TPF_{i-1})^2}$.)

A similar procedure was performed for the proper binormal model. We calculated the conventional AUC for each of 250 values of c distributed uniformly between -1 and 1, and (at each such value of c) for 250 values of d_a uniformly distributed between 0 and 4. For each of these 62,500 pairs of parameter values, we also calculated the corresponding value of $\overline{C}_\sigma$ (again using the approximation for arc length described for the conventional model).

4. RESULTS

The calculated values of A_z and of $\overline{C}_\sigma$ for the conventional binormal model are shown in isopleth ("contour") plots in Fig. 1. Similarly, the calculated values of the conventional AUC and $\overline{C}_\sigma$ for the proper binormal model are shown in isopleth plots in Fig. 2.

Although difficult to discern from the plot, the isopleths in Fig. 1 do in fact cross, particularly in the lower left region. For example, the parameter pair $(a = 0.4819, b = 0.5060)$ corresponds to an A_z value of 0.6663 and

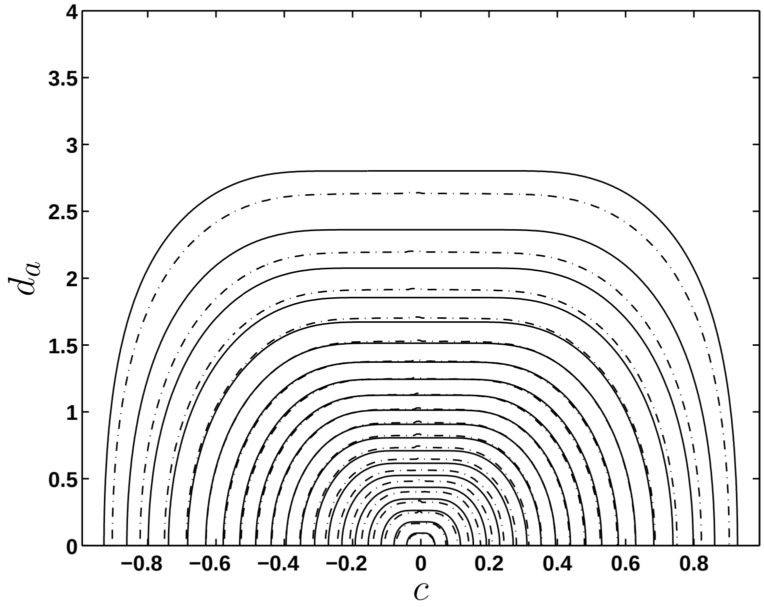


Figure 2. Isopleths of the conventional AUC performance metric (solid lines) and of the proposed $\overline{C}_\sigma$ metric (dash-dotted lines), for various values of the c and d_a parameters of the proper binormal model.

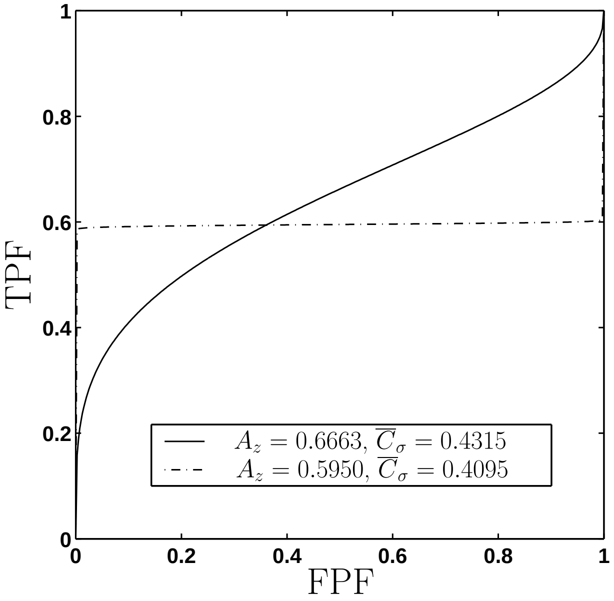


Figure 3. ROC curves generated under the conventional binormal model with parameter values of $(a = 0.4189, b = 0.5060)$ (solid curve), and $(a = 0.2410, b = 0.0080)$ (dash-dotted curve).

a $\overline{C}_\sigma$ of 0.4315, while the parameter pair $(a = 0.2410, b = 0.0080)$ corresponds to an ROC curve which has both a lower A_z of 0.5950 and a lower $\overline{C}_\sigma$ of 0.4095. (Recall that, as its name implies, the SAEC $\overline{C}_\sigma$ is a "cost", and thus lower values are intended to be "preferable," in contrast to A_z and the conventional AUC.) These two ROC curves are plotted (conventionally, using TPF as the ordinate) in Fig. 3.

5. DISCUSSION

It is evident from Fig. 1 that the proposed performance metric $\overline{C}_\sigma$ does not perform identically to the conventional AUC in all situations (*i.e.*, for arbitrary decision rules). This is illustrated in more detail in Fig. 3; if the two curves represented observers (radiologists or imaging systems, for example) which one wished to rank in order of performance, then the two performance metrics would disagree as to which were actually preferable. This is understandable given the shapes of the curves; the system with slightly lower A_z is so severely "hooked" that its arc length will be very close to two, driving down the "cost" $\overline{C}_\sigma$ to a greater extent than the loss in conventional AUC.

It should be recalled, however, that in practical situations in which such a severe "hook" is seen in the ROC curve, the observational data themselves do not usually support such a fitting of the curve.[6] Even aside from such data sampling and curve-fitting issues, comparing two systems when at least one of them has an ROC curve with such a large "hook" is often problematic (compare the well-known situation when two systems have very similar AUCs, but "cross," making the decision of which system to prefer dependent on the region of ROC space in which one chooses to operate). In short, the fact that $\overline{C}_\sigma$ does not agree exactly with a performance metric such as A_z, itself known to be imperfect, is not necessarily a fatal flaw.

The results presented in Fig. 2 are far more surprising. There appear to be no visible "crossings" of the isopleths for any choices of parameters c and d_a. Although this result still needs to be confirmed analytically, it would if found true imply that $\overline{C}_\sigma$ and the conventional AUC under the proper binormal model are equivalent performance metrics. Whether this equivalence could be extended to arbitrary ideal observer models (*i.e.*, those for arbitrary PDFs rather than the binormal model) would also be an important area for further investigation.

The extensibility of the proposed performance metric to tasks with more than two classes is quite plausible, but much remains to be done here as well. Preliminary work in this direction suggests that it may be possible to apply an extension of L'Hôpital's rule to the integrals in Eq. 15 in the situation where they approach 0 due to dimensionality considerations. However, the resulting limit appears to depend strongly on the underlying data PDFs (a counterintuitive result given the behavior of two-class near-guessing observers, whose ROC curves all approach the diagonal line regardless of the data PDFs). More careful work will be necessary to validate or refute these claims.

Related to the issue of dimensionality just mentioned is the situation of the "discrete" observer, *i.e.*, an observer which operates only at discrete operating points in ROC space (this applies to the two-class observer as well as those with more classes). We have so far been unable to usefully generalize the definition of $\hat{\gamma}_R$ and thus Eq. 15 to this situation, even in the two-class case. It remains to be seen whether this last issue is an important one or not.

6. CONCLUSIONS

We have proposed a novel ROC performance metric, the SAEC. Although grounded in the same theoretical framework as the expected utility of the ideal observer, its practical realization involves readily comprehensible quantities — the AUC and the arc length along the ROC curve in a two-class task, and the surface-averaged integral of a well-defined scalar in a task with more than two classes.

Although the properties of this performance metric have yet to be thoroughly investigated, preliminary results are quite encouraging. We have high hopes that this performance metric will allow comparison of observers in classification tasks of varying complexity, without suffering from the drawbacks that other performance metrics, such as the HUH, have been shown to possess.

ACKNOWLEDGMENTS

This work was supported by grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D. C. Edwards, principal investigator). Charles E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

REFERENCES

1. J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
2. C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine* **VIII**(4), pp. 283–298, 1978.
3. D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.* **31**, pp. 81–90, 2004.
4. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
5. D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.
6. C. E. Metz and X. Pan, "'Proper' binormal ROC curves: Theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**, pp. 1–33, 1999.
7. D. C. Edwards, C. E. Metz, and R. M. Nishikawa, "The hypervolume under the ROC hypersurface of 'near-guessing' and 'near-perfect' observers in N-class classification tasks," *IEEE Trans. Med. Imag.* **24**, pp. 293–299, 2005.
8. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., New York, 1991.
9. D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.* **50**, pp. 478–487, 2006.
10. D. C. Edwards and C. E. Metz, "Optimization of an ROC hypersurface constructed only from an observer's within-class sensitivities," in Proc. SPIE Vol. 6146 *Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment*, Yulei Jiang and Miguel P. Eckstein, eds., pp. 61460A1–61460A7, (SPIE, Bellingham, WA), 2006.
11. D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.* , 2006. (submitted).
12. S. I. Grossman, *Multivariable Calculus, Linear Algebra, and Differential Equations: Second Edition*, Harcourt Brace Jovanovich, San Diego, CA, 1986.
13. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statist. Med.* **17**, pp. 1033–1053, 1998.